UNIVERSIDAD DE CÓRDOBA

E.T.S.I.A.M.

DEPARTAMENTO DE GENÉTICA

TESIS DOCTORAL

«Caracterización de la variación para los loci *Glu-3* y *Ha* en especies diploides de los géneros *Aegilops* y *Triticum*»

Doctoranda:

Susana Cuesta Ureña

Directores:

Dr. Juan Bautista Álvarez Cabello

Dr. Carlos Guzmán García

Octubre 2015

TITULO: *Caracterización de la variación para los loci Glu-3 y Ha en especies diploides de los géneros Aegilops y Triticum.*

AUTOR: *Susana Cuesta Ureña*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TÍTULO DE LA TESIS: CARACTERIZACIÓN DE LA VARIACIÓN PARA LOS LOCI *GLU-3* Y *HA* EN ESPECIES DIPLOIDES DE LOS GÉNEROS *AEGILOPS* Y *TRITICUM*

DOCTORANDO/A: SUSANA CUESTA UREÑA

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

La Tesis ha presentado un desarrollo coherente con las previsiones realizadas en su momento. Dentro de la misma se ha realizado una estancia breve en el Department of Plant Sciences and Plant Pathology en Montana State University, trabajando en las investigaciones llevadas a cabo por el profesor Michael Giroux.

Los trabajos desarrollados y que conforman esta Tesis han permitido la detección de nuevas variantes alélicas para los loci motivo de estudio, los cuales han sido caracterizados a través del análisis de sus secuencias nucleotídicas. Dado el papel que estas proteínas tienen en parámetros de la calidad del trigo como la dureza (puroindolinas y Gsp-1 sintetizadas en el locu *Ha*) y fuerza del gluten (LMWGs sintetizadas en los loci *Glu-3*), junto con el hecho de que las especies evaluadas están relacionadas con los genomas presentes en el trigo moderno, hace que la información generada pueda ser utilizada en el medio-largo plazo para el desarrollo de nuevos cultivares de trigo que presenten nuevas propiedades para los dos parámetros citados.

En base a este desarrollo, se ha optado por la presentación de esta Tesis como compendio de publicaciones, comprendiendo un total de 5 publicaciones (capítulos I al V del documento), de los cuales actualmente se encuentran publicadas una en la revista «*Journal of Experimental Botany*» y otra en la revista «*Journal of Cereal Science* », junto con un tercero en prensa en la revista «*Theoretical and Applied Genetics*», todas ellas incluidas en algunas de las categoría del «*Science Citation Index*». En cuanto a las otras dos, una de ellas está en revisión en la revista « *Theoretical and Applied Genetics*», y la otra está enviada a «*Molecular Breeding*», revistas todas ellas incluidas en el «*Science Citation Index*».

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 16 de octubre de 2015

Firma del/de los director/es

Fdo.: Juan Bautista Alvarez Cabello

Fdo.: Carlos Guzmán García

Juan Bautista Álvarez Cabello, Profesor Titular de la Universidad de Córdoba y Carlos Guzmán García, Laboratorio de Química y Calidad de Trigo, Programa Global de Trigo, Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), Texococo, México.

INFORMAN:

Que el trabajo titulado: **«Caracterización de la variación para los loci *Glu-3* y *Ha* en especies diploides de los géneros *Aegilops* y *Triticum*»**, realizado por D[a]. Susana Cuesta Ureña, bajo su dirección, se considera ya finalizado y puede ser presentado para su exposición y defensa como Tesis Doctoral en el Departamento de Genética de la Universidad de Córdoba.

Firmado en Córdoba, a 19 de octubre de 2015

Fdo. Juan Bautista Álvarez Cabello

Fdo. Carlos Guzmán García

Esta investigación ha sido financiada por los Proyectos AGL2010-19643-C02-01 y AGL2014-52445-R del Ministerio de Economía y Competitividad, cofinanciada con el Fondo Europeo para el Desarrollo Regional (FEDER) de la Unión Europea, así como por el Proyecto P11-AGR-7920 de la Junta de Andalucía.

La doctoranda agradece, asimismo, al Ministerio de Economía y Competitividad (Programa FPI) y Fondo Social Europeo por la concesión de una beca predoctoral.

A mi madre y a mi pareja

AGRADECIMIENTOS

Quisiera expresar mi más sincero agradecimiento a todas las personas que han contribuido de algún modo a la realización de este trabajo.

En primer lugar, a mi director Juan Bautista Álvarez por ofrecerme la oportunidad de realizar este doctorado, por su gran dedicación, guía y disponibilidad para enseñarme todo lo necesario en estos años. Gracias también a mi otro director Carlos Guzmán García por su ayuda.

A todos los miembros del Departamento de Genética de la Universidad de Córdoba. A todos los profesores, trabajadores, becarios y alumnos que en algún momento me han ayudado con este trabajo. A Marcela, mi compañera de grupo, por toda la ayuda y por aguantarme todos estos años, Latifeh y Gabi, por todos los favores que me hicieron en el laboratorio y por su ayuda, Tere, Francisco y demás compañeros, por su compañía, y a todos los demás que he conocido en Córdoba, los que ya se fueron y los que aún siguen aquí, como Mennat.

A mis amigos de Badalona, por hacer que desconectara cuando iba a mi tierra. A los de la uni de Barna, Ana, Lara y Javi, por venir a visitarme y hacer más amena mi estancia en Córdoba.

A toda mi familia, especialmente a mi madre por apoyarme y creer constantemente en mí durante mi doctorado y todos mis estudios. Además, a mi hermana Inés por animarme y ayudarme durante estos años y toda mi vida.

Y por último, y más importante, a mi pareja, por su paciencia, ánimo y apoyo constante durante todos estos años. Sin ti no hubiera sido capaz.

LISTADO DE PUBLICACIONES OBTENIDAS Y ESTATUS DE LAS MISMAS A LA FECHA DE FIRMA DEL PRESENTE DOCUMENTO

1. S. Cuesta, C. Guzmán, J.B. Alvarez (2013) Allelic diversity and molecular characterization of *Puroindoline* genes in five diploid species of the *Aegilops* genus. *Journal of Experimental Botany* **64**: 5133-5143.

2. S. Cuesta, J.B. Alvarez, C. Guzmán (2015) Characterization and sequence diversity of the *Gsp-1* gene in diploid species of the *Aegilops* genus. *Journal of Cereal Science* **63**: 1-7.

3. S. Cuesta, C. Guzmán, J.B. Alvarez (2015) Molecular characterization of novel LMW-m and -s genes from four *Aegilops* species (*Sitopsis* section) and comparison with those from the *Glu-B3* locus of common wheat (under review, *Molecular Breeding*).

4. S. Cuesta, C. Guzmán, J.B. Alvarez (2015) Molecular characterization of novel LMW-i glutenin subunit genes from *Triticum urartu* Thum. ex Gandil. *Theoretical and Applied Genetics* **128**: 2155-2165.

5. S. Cuesta, J.B. Alvarez, C. Guzmán (2015) Identification and molecular characterization of novel LMW-m and -s glutenin genes, and a chimeric -m/-i glutenin gene in three diploid *Triticum* species (under review, *Molecular Breeding*).

Nota: A fin de establecer una coherencia formal a lo largo del presente documento, se han uniformizado las referencias y se han editado los trabajos originales, eliminando de los mismos el apartado de referencias, el cual ha sido agrupado al final del documento.

ÍNDICE

RESUMEN

El trigo es uno de los cultivos más importantes a nivel mundial. La búsqueda de nueva variabilidad que pueda ser utilizada en programas de mejora para aumentar la base genética del trigo moderno ha sido un tema principal de estudio en las últimas décadas. Consecuentemente, la caracterización y evaluación de los recursos genéticos disponibles es esencial y en los últimos años se ha extendido a especies relacionadas con los genomas del trigo. Entre estas especies se encuentran especies del género *Triticum* y *Aegilops*.

El objetivo de esta Tesis Doctoral ha sido la evaluación y caracterización de la variación alélica detectada para los genes *Pin* y *Gsp-1*, así como para los genes de LMWG, en especies diploides de los géneros *Triticum* y *Aegilops*. En el caso de las especies del género *Triticum*, dado que los genes *Pin* habían sido previamente analizados en otro trabajo previo del grupo, se procedió directamente a la caracterización molecular de los genes de LMWGs.

Para los genes *Pin*, una colección de 82 accesiones de *Aegilops* fue evaluada en función de la movilidad electroforética de sus amplicones y posteriormente se secuenciaron las principales variantes alélicas. La combinación de ambos genes *Pin* reveló 42 genotipos, cuyas respectivas accesiones fueron adicionalmente evaluadas para *Gsp-1* siguiendo el mismo procedimiento. En ambos casos se detectaron nuevas variantes que podrían afectar a la dureza del grano. La más significativa fue una mutación encontrada en el codón de inicio de una variante de PINA (ATG por TTG) que podría generar una reducción en la expresión de la proteína resultando en una dureza de grano intermedia entre *soft* y *hard*.

Los genes de LMWG fueron caracterizados para la sección *Sitopsis* del género *Aegilops*, y las especies diploides de *Triticum*. Entre las variantes detectadas se encontró una con características combinadas de un gen de LMW-m y un gen de LMW-i (LMW-m/-i) que podría tener un efecto distintivo sobre calidad de la masa. Además, el análisis de los epítopos reactivos para la enfermedad celíaca indicó que las subunidades LMW-i de *T. urartu* fueron las más reactivas, mientras que las LMW-s de la sección *Sitopsis* fueron las menos reactivas.

Finalmente, el estudio reveló que los genes *Gsp-1* son poco eficaces para estudios filogenéticos debido a su reducido tamaño; por el contrario, los genes de LMWG revelaron nuevos conocimientos en la evolución y composición de esta familia multigénica. Del mismo modo, los genes de LMWG permitieron estudiar las relaciones

filogenéticas entre las especies estudiadas, corroborando la hipótesis de que el genoma B de los trigos poliploides podría tener un origen polifilético.

En conclusión, los resultados sugieren que las especies evaluadas podrían constituir una importante fuente de variabilidad genética para los genes estudiados y ser utilizadas en programas de mejora para el desarrollo de nuevas variedades de trigo que presenten nuevas propiedades viscoelásticas y nuevos rangos de dureza, incluido un potencial uso en la elaboración de productos aptos para celíacos. Secundariamente, el análisis filogenético resulta ser una buena herramienta para el estudio de la evolución y composición de la familia multigénica de genes LMWG. Como consecuencia de todo lo anterior, es esencial el mantenimiento de estas especies con el fin de salvaguardar su diversidad para que puedan ser utilizadas como recursos genéticos en programas de mejora del trigo actual.

ABSTRACT

Wheat is one of crops most important worldwide. The search of novel variability that can be used in breeding programmes to increase genetic base of modern wheat has been a major subject of study in the last decades. Consequently, the characterization and evaluation of genetic resources available is essential, and in the last years has been extended to related species with wheat genomes. Among this species are the species of *Triticum* and *Aegilops* genus.

The aim of this PhD Thesis was the evaluation and characterization of allelic variation detected for *Pin* and *Gsp*-1 genes, as well as LMWG genes, in diploid species of *Triticum* and *Aegilops* genus. In the case of the *Triticum* genus species, since the *Pin* genes had been previously analysed in a previous work of the group, it proceeded directly to the molecular characterization of LMWG genes.

For the *Pin* genes, a collection of 82 *Aegilops* accessions was evaluated according to their amplicons electrophoretic mobility and subsequently the main allelic variants were sequenced. The combination of both *Pin* genes revealed 42 genotypes, whose respective accessions were further evaluated for *Gsp-1* following the same procedure. In both cases novel variants that could affect to grain hardness were detected. The most significant was a mutation found in start codon of PINA variant (ATG to TTG) that could lead to reduction in protein expression resulting in a grain hardness intermediate between soft and hard.

LMWG genes were characterized in the *Sitopsis* section of *Aegilops* genus, and the diploid species of *Triticum*. Among the variants detected, one showed combined characteristics of a LMW-m gene and LMW-i gene (LMW-m/-i) that could have a distinctive effect on dough quality. Furthermore, analysis of reactive epitopes for celiac disease indicated that LMW-i subunits of *T. urartu* were more reactive, while the LMW-s subunits of *Sitopsis* section were less reactive.

Finally, the study revealed that the *Gsp-1* genes are ineffective for phylogenetic studies because of their reduced size; conversely, LMWG genes revealed novel knowledge in the evolution and composition of this multigene family. Similarly, LMWG genes allowed to study the phylogenetic relationships among species studied, confirming the hypothesis that B genome polypoid wheats could have a polyphyletic origin.

In conclusion, the results suggest that the species evaluated could be an important source of genetic variability for the genes studied and be used in breeding

programs for developing novel wheat varieties that present novel viscoelastic properties and novel hardness ranges, including a potential use in the elaboration of products suitable for celiac patients. Secondly, the phylogenetic analysis results to be a good tool for the study of the evolution and composition of multigene family of LMWG genes. As result of the above, the maintenance of these species in order to safeguard their diversity so that they can be used as resources genetic in breeding programs of modern wheat is essential.

INTRODUCCIÓN GENERAL

Actualmente, de acuerdo a su producción, el trigo es el cuarto mayor cultivo a nivel mundial con casi 716 millones de toneladas y una superficie cultivada de 219 millones de hectáreas, representando casi el 26% de la producción total de los cereales a nivel mundial, según datos consolidados de FAO de 2013 (Tabla 1).

Tabla 1. Datos de producción y superficie cultivada de los principales cereales.

Cultivo	Producción (Tm)	%	Superficie (Ha)	%
Maíz	1.018.111.958	36,6	185.121.343	25,7
Arroz	740.902.532	26,7	165.163.423	22,9
Trigo	715.909.258	25,8	219.046.706	30,4
Cebada	143.959.778	5,18	49.148.479	6,81
Sorgo	62.295.137	2,24	42.227.048	5,85
Mijo	29.864.147	1,07	33.118.792	4,59
Avena	23.880.997	0,86	9.779.904	1,35
Centeno	16.686.795	0,6	5.760.278	0,8

Datos según FAO (2013).

Los principales países productores de trigo son: China, India, los Estados Unidos de América y la Federación Rusa, con algo más del 45 % de la producción mundial (Tabla 2). España se sitúa en posición decimoctava con una producción de 7,6 millones de toneladas y una superficie cultivada de 2,1 millones de hectáreas, que representa el 1% de la producción mundial (Tabla 2).

Tabla 2. Datos de producción y superficie cultivada de trigo.

País	Producción (Tm)	%	Superficie (Ha)	%
China	121.956.400	17,0	24.119.335	11,0
India	93.510.000	13,1	29.650.000	13,5
Estados Unidos de América	57.966.658	8,1	18.274.206	8,3
Federación Rusa	52.090.797	7,3	23.371.410	10,7
Francia	38.613.900	5,4	5.323.000	2,4
Canadá	35.529.600	5,0	10.441.500	4,8
Alemania	25.019.100	3,5	25.019.100	11,4
Paquistán	24.211.400	3,4	8.686.602	4,0
Australia	22.855.576	3,2	12.979.382	5,9
Ucrania	22.793.000	3,2	6.566.000	3,0
Resto del mundo	213.760.227	29,9	52.494.271	24,0
España	7.602.600	1,1	2.121.900	1,0

Datos según FAO (2013).

El trigo es el cultivo dominante en las zonas de clima templado, siendo no obstante cultivado en otras zonas climáticas del mundo. La gran expansión del cultivo es debido a su gran adaptación a diferentes ambientes, a su fácil cosecha, almacenamiento y conservación, y a las peculiares propiedades de la masa obtenida con su harina que permite la elaboración de una extensa variedad de productos (Shewry 2009). El 94% del trigo cultivado a nivel mundial es trigo harinero, mientras que el 6% restante corresponde principalmente a trigo duro. Pequeñas cantidades de otras especies de trigo (escaña, emmer y espelta) son también cultivadas.

Origen del trigo

La Agricultura se desarrolló durante la Revolución Neolítica, provocando una profunda transformación en la historia humana que pasó de sociedades nómadas de cazadores-recolectores a sociedades sedentarias de ganaderos-agricultores. Los hallazgos arqueológicos sugieren que los primeros asentamientos agrícolas se desarrollaron en el Próximo Oriente en torno al 10.500 a.C., en la zona conocida como Creciente Fértil. Posteriormente, la Agricultura se desarrollaría en el Extremo Oriente (valle del río Yang-Tse) entre el 6.500-5.5000 a.C., aparenciendo entre el 5.000-4.000 a.C. en América Central y la zona Andina (Harlan 1992).

Entre las primeras plantas en ser cultivadas destacan tanto los cereales como las leguminosas, debido a su alto contenido calórico y su fácil conservación. De hecho, en el Creciente Fértil, dos de los cultivos más antiguos fueron la cebada y el trigo, cereales que fueron, y en el caso del trigo siguen siendo, la base alimentaria en la mayor parte del Mundo. Así, Asia basó su alimentación en el arroz, África en el sorgo, América en el maíz y la Cuenca del Mediterráneo en el trigo.

El término trigo se aplica en sentido extenso a las especies del género *Triticum* (familia *Poaceae*, subfamilia *Pooideae*, tribu Triticeae). Bajo esta denominación se encuentra un complejo poliploide formando por especies silvestres y cultivadas, agrupadas en tres niveles diferentes de ploidía:

- Diploides: presentan un solo genoma (**A**) constituido por 7 parejas de cromosomas homólogos (2n = 2× = 14).
- Tetraploide: presentan dos genomas diferentes (**AB**), con 14 parejas de cromosomas (2n = 4× = 28).
- Hexaploide: presentan tres genomas diferentes (**ABD**) y 21 parejas de cromosomas (2n = 6× = 42).

La primera especie del género *Triticum* cultivada por el hombre fue el trigo einkorn o escaña (*Triticum monococcum* L. *ssp. monococcum*, $\mathbf{A^mA^m}$), procedente de la domesticación de *T. monococcum* ssp. *aegilopoides* Link em. Thell (syn. *T. boeoticum* Boiss.), proceso que se comenzó a realizar en la zona del Creciente Fértil (Fig. 1). Otra especie diploide de este género que tiene gran importancia es *T. urartu* Thum. ex Gandil ($\mathbf{A^uA^u}$), especie silvestre de la cual procede el genoma **A** de los trigos poliploides (Dvorak et al. 1993; Matsuoko 2011).

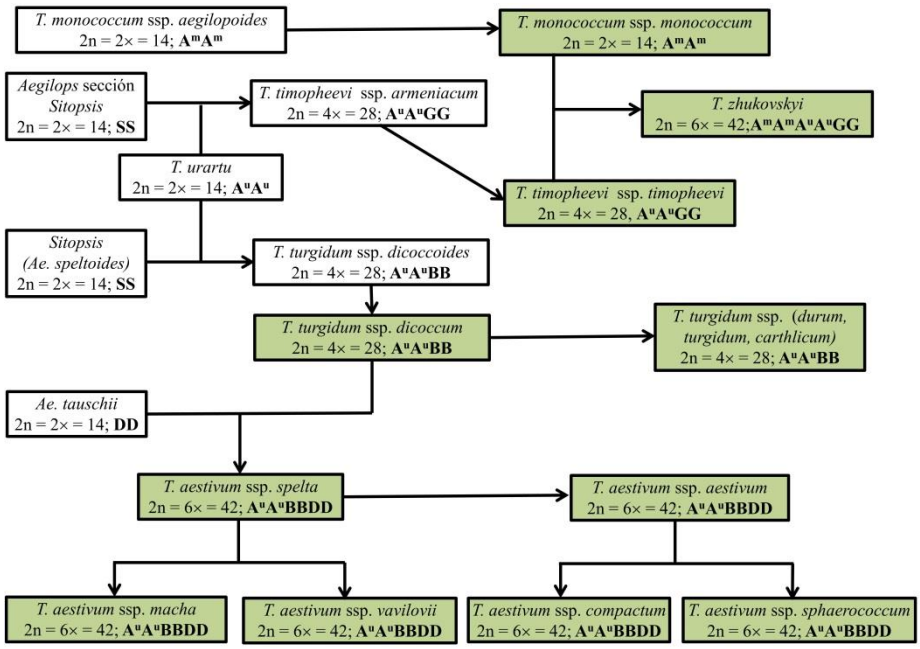


Figura 1. Esquema de la evolución de las diferentes especies de trigo. Las especies en verde son o han sido cultivadas en algún momento.

Las especies tetraploides fueron originadas a partir de cruzamientos espontáneos de especies del género *Aegilops* con *T. urartu* en, al menos, dos diferentes eventos evolutivos (Fig. 1). Por un lado, se desarrolló el emmer silvestre (*T. turgidum* ssp. *dicoccoides* Körn. ex Asch. & Graebner em. Thell.; $\mathbf{A^uA^uBB}$), muy probablemente como consecuencia del cruzamiento con *Ae. speltoides* Tausch (2n = 2× = 14, **SS**) en el Creciente Fértil (Petersen et al. 2006). No obstante, el origen del genoma B aún suscita gran controversia y las otras cuatro especies de *Aegilops* pertenecientes, al igual que *Ae. speltoides*, a la sección *Sitopsis* han sido también consideradas como posibles donadoras (Haider et al. 2013). A este respecto, Sarkar y Stebbins (1956) sugirieron la posibilidad

de un origen polifilético de este genoma, en el cual estarían implicadas todas estas especies en otros tantos eventos evolutivos. A partir de la domesticación del emmer silvestre se originó el emmer cultivado o escanda menor (*T. turgidum* ssp. *dicoccum* Schrank em. Thell.), desde el cual evolucionaron el resto de trigos tetraploides que hoy conocemos, entre los que destaca el trigo duro (*T. turgidum* ssp. *durum* Desf. em. Husn.). Por otro lado, estas mismas especies (*Ae. speltoides* y *T. urartu*) se vieron involucradas en un evento local en el occidente de Georgia (Kilian et al. 2007), generando otra especie tetraploide (*T. timophevii* ssp. *armeniacum* Jakubz. em. Slageren; **A^uA^uGG)**, cuya especie cultivada es *T. timophevii* ssp. *timophevii*.

A partir de esta última especie y su cruzamiento con la escaña cultivada, la cual aportó otro genoma A a la especie, se originó *T. zhukovskyi* Menabde & Ericz. (**$A^mA^mA^uA^uGG$)** (Baum y Bailey 2004), un trigo hexaploide al margen del gran grupo de los trigos hexaploides generados por el cruzamiento y posterior duplicación entre el trigo emmer cultivado y *Ae. tauschii* Coss. (2n = 2× = 14, **DD**), que dio lugar al trigo espelta (*T. aestivum* ssp. *spelta* L. em. Thell., **A^uA^uBBDD**). Este evento tuvo lugar probablemente en algún campo de cultivo de la zona del Mar Caspio, dentro del área de crecimiento natural de *Ae. tauschii* (Dvorak et al. 1998). A partir de la evolución del espelta se originó el trigo harinero (*T. aestivum* ssp. *aestivum* L. em. Thell.) - Fig.1-.

Como previamente se ha mencionado, el género *Aegilops* es importante en la evolución del trigo, ya que algunas de sus especies han contribuido en el origen de los genomas del trigo. Este género, al igual que el del trigo, pertenece a la tribu Triticeae y está compuesto por 23 especies, de las cuales 11 son diploides y 12 son alopoliploides (10 tetraploides y 2 hexaploides) y son divididas dentro de cinco secciones, junto con el subgénero *Ambylopyrum* formado por la antigua especie *Ae. mutica* (van Slageren 1994) -Tabla 3-:

- Sect. *Aegilops* L.;
- Sect. *Comopyrum* (Jaub. & Spach) Zhuk.;
- Sect. *Cylindropyrum* (Jaub. & Spach) Zhuk.;
- Sect. *Sitopsis* (Jaub. & Spach) Zhuk.,
- Sect. *Vertebrata* Zhuk. emend. Kihara

Tabla 3. Clasificación de las especies del género *Aegilops* según van Slageren (1994).

Especie	Genoma
Sección *Aegilops* L.	
Ae. biuncialis Vis.	UM
Ae. geniculata Roth	UM
Ae. neglecta Req. ex Bertol.	UM/UMN
Ae. peregrina (Hack. in J.Fraser) Marie & Weiller	SU
var. *peregrina*	
var. *brachyathera* (Boiss.) Marie & Weiller	
Ae. umbellulata Zhuk.	U
Ae. columnaris Zhuk.	UM
Ae. triuncialis L.	UC
var. *triuncialis*	
var. *persica* (Boiss.) Eig	
Ae. kotschyi Boiss	US
Sección *Comopyrum* (Jaub. & Spach) Zhuk.	
Ae. comosa Sm. in Sibth. & Sm.	M
var. *comosa*	
var. *subventricosa* Boiss.	
Ae. uniaristata Vis.	N
Sección *Cylindropyrum* (Jaub. & Spach) Zhuk.	
Ae. caudata L. (*Ae. markgrafii* (Greuter) Hammer)	C
Ae. cylindrica Host	CD
Sección *Sitopsis* (Jaub. & Spach) Zhuk.	
Ae. bicornis (Forssk.) Jaub. & Spach	S^b
var. *bicornis*	
var. a*nathera* Eig	
Ae. longissima Schweinf. & Muschl.	S^l
Ae. searsii Feldman & Kislev ex Hammer	S^s
Ae. sharonensis Eig	S^{sh}
Ae. speltoides Tausch	S
var. *speltoides*	
var. *ligustica* (Savign.) Fiori	
Sección *Vertebrata* Zhuk. emend. Kihara	
Ae. juvenalis (Thell). Eig	DMU
Ae. tauschii Coss.	D
Ae. vavilovii (Zhuk.) Chennav.	DMS
Ae. crassa Boiss.	DM/DDM
Ae. ventricosa Tausch	DN
***Ambylopyrum* (Jaub. & Spach) Eig**	
Ambylopyrum muticum (Boiss.) Eig	T
var. *muticum*	
var. *loliaceum* (Jaub. & Spach) Eig	

Los usos del trigo

El trigo es usado para la elaboración de diferentes productos, siendo el pan, el más importante y conocido desde la Antigüedad. Los primeros documentos escritos que hacen referencia al pan datan del año 2600 a.C., aunque se han encontrado hallazgos

arqueológicos más antiguos en Babilonia, en torno al 4000 a.C. Los primeros en utilizar la levadura para la fermentación y esponjamiento de la masa en la panificación fueron los egipcios. El pan se utilizaba como un símbolo del estatus social, así las clases altas tomaban un pan realizado con harina de alta calidad, mientras que las clases bajas comían un pan tosco y de baja calidad (Harlan 1981). El conocimiento sobre la panificación pasó de los egipcios a los griegos, aunque estos preferían tomar la harina en forma de gachas. Posteriormente pasó a los romanos, para los cuales tuvo una gran importancia, desarrollando su producción a nivel industrial en la cuenca del Mediterráneo y dándole valor desde el punto de vista cultural y religioso. Al igual que los egipcios utilizaron el pan como símbolo de la posición social que ocupaban.

Con la llegada de la Revolución Industrial, los procesos de elaboración del pan fueron mecanizados, lo que permitió el desarrollo de nuevos y diferentes productos, usando diferentes materias primas o incluso otros cereales. Debido al aumento demográfico, nuevas variedades de trigo con mayor rendimiento y aptas para su procesamiento mecánico fueron desarrolladas.

El trigo también es utilizado para la elaboración de pasta y couscous, a partir de sémola de trigo duro, así como para la elaboración de cerveza. Actualmente el trigo también es usado para la industria no alimentaria, donde destaca la industria cosmética e incluso la utilización para la producción de bioetanol o biomasa (Bell et al. 1995).

La calidad del trigo

La calidad del trigo depende del producto que se vaya a elaborar, ya que una variedad de trigo con propiedades adecuadas para la elaboración de un producto determinado puede ser inadecuada para la elaboración de otros productos. Por tanto, la calidad de la harina se podría definir como la capacidad de una determinada variedad de trigo para producir un producto específico.

La calidad del trigo es un carácter altamente complejo influido por factores ambientales, tales como el suelo, el clima, las técnicas de cultivo y almacenamiento, además de poseer un importante componente genético. Tres son los grupos de proteínas presentes en el grano que juegan un papel esencial en la calidad:

- las proteínas asociadas a la dureza del grano;
- las proteínas de reserva del endospermo; y
- las proteínas sintetizadoras del almidón.

Proteínas asociadas a la dureza

La dureza o textura del endospermo del grano es una importante propiedad que afecta a la calidad del trigo. Es definida como el grado de adhesión entre las proteínas de la matriz y los gránulos de almidón del endospermo. De acuerdo a esta característica, el trigo harinero es clasificado como blando (*soft*) y tenaz o duro (*hard*) y, como consecuencia, determina el uso final del mismo (Morris y Rose 1996).

En los trigos *hard*, el grado de adhesión entre las proteínas de la matriz y los gránulos de almidón es mayor que para los trigos *soft*. La resistencia, energía y tiempo requerido en la molienda de los trigos *hard* son más altos que para los trigos *soft*. La molienda de los trigos *hard* produce harinas con partículas de tamaño grande y con muchos gránulos de almidón dañados. La gran cantidad de almidón dañado presente en la harina es esencial para la absorción de agua de la misma, lo que las hace idóneas para la fabricación de pan. Sin embargo, los granos de trigo *soft* se rompen más fácilmente, produciendo harinas con partículas más finas y con menor cantidad de almidón dañado, por lo que son idóneas para la elaboración de galletas y pasteles (Morris y Rose 1996).

La dureza ha sido asociada a un complejo de proteínas llamado friabilina, formado por dos componentes principales: puroindolina a (PINA) y puroindolina b (PINB), junto con un componente minoritario, la proteína de la suavidad del grano (GSP-1, *Grain Softness Protein*) (Morris 2002). Los genes sintetizadores de estas proteínas, denominados respectivamente *Pina*, *Pinb* y *Gsp*, están localizados en el locus *Ha* (*Hardness*) situado en el brazo corto del cromosoma 5 de todas las especies de la familia *Poaceae* (Wilkinson et al. 2013).

La dureza ha sido relacionada con la presencia/ausencia o modificaciones de algunos de los genes *Pin* (Bhave y Morris 2008a; Feiz et al. 2009). El caso extremo se produce en el trigo duro, donde existe una delección parcial en los loci ortólogos de los cromosomas 5A y 5B, acontecida durante la generación de los trigos tetraploides, que comprendía los genes *Pin* de estos cromosomas (Fig. 2), lo que origina que ambas puroindolinas (PINs) estén ausentes y la textura sea muy dura o extradura. Estos genes se recuperaron en el trigo harinero a través del aporte del cromosoma 5D derivado de *Ae. tauschii* (Li et al. 2008a). En trigo harinero, la presencia de ambas PINs en su forma funcional está asociada a la textura blanda de los trigos *soft*; mientras que los trigos *hard* poseen texturas duras debido a la ausencia de una de las dos PINs o a algún tipo de alteración en estos genes que generen proteínas no completamente funcionales (Bhave y Morris 2008b).

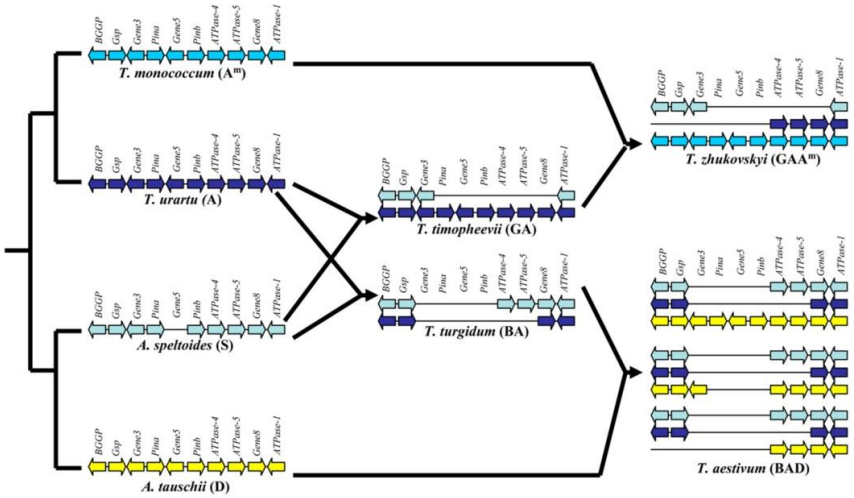


Figura 2. Esquema y evolución del locus *Ha* en trigo (Li et al. 2008a).

Las PINs son proteínas básicas con un peso molecular en torno a los 13 kDa. Poseen un esqueleto formado por 10 residuos de cisteína altamente conservado y un dominio hidrofóbico rico en triptófano que actúa como sitio de unión con los lípidos polares presentes en la membrana de los amiloplastos (Pauly et al. 2013). Esta capacidad de unión a lípidos es la que determina su papel en la dureza del grano. Recientemente, Gedye et al. (2004) sugirieron un posible papel para la GSP-1 en la dureza del grano, aunque otros estudios habían indicado previamente lo contrario (Tranquilli et al. 2002). Su similitud con las PINs, incluyendo la conservación del esqueleto de cisteína y el dominio triptófano, sugiere que podría interactuar de la misma manera con los gránulos de almidón y tener un papel secundario en la dureza. No obstante, su papel en la dureza sigue suscitando cierta controversia. Recientemente, Elmorjani et al. (2013) aislaron la GSP-1 desde un trigo *soft* e indicaron que no tenía capacidad de interacción con lípidos; sin embargo sólo una isoforma fue estudiada (*Gsp-1b*). La principal diferencia con las PINs es la presencia de un dominio de 15 residuos en el N-terminal asociado al péptido arabinogalactano (AGP), el cual forma parte de los polisacáridos no almidonados sintetizados por las células de la pared del endospermo (Van den Bulck et al. 2002). Estos AGP se localizan en las membranas plasmáticas, incluyendo esas de los amiloplastos, de manera que podrían influir en la fuerza de adhesión entre los gránulos de almidón y la matriz proteica. El estudio de Bettge y Morris (2000) indicó que hasta un 76% de la variación de la dureza entre trigos *soft* podría ser causada por polisacáridos no almidonados.

La búsqueda de nuevos alelos de estos genes es importante para aumentar el rango de texturas disponibles y mejorar la calidad del trigo actual (Bhave y Morris 2008a). En los últimos años, la búsqueda de nuevas variantes se ha extendido a especies relacionadas y un amplio rango de durezas ha sido detectado (Gautier et al. 2000; Lillemo et al. 2002; Simeone et al. 2006; Guzmán et al. 2011). Algunas de las mutaciones encontradas provocan que el grano sea incluso más blando que el encontrado en trigo harinero (Gedye et al. 2004; Feiz et al. 2009). Existen otros factores bioquímicos, así como otros loci que podrían tener un menor papel en este carácter (Giroux et al. 2000; Bhave y Morris 2008). Sin embargo, el locus *Ha* sigue siendo el principal responsable de la dureza, explicando los genes *Pin* más del 60% de la variación en este carácter (Campbell et al. 1999).

Proteínas de reserva

Las prolaminas o proteínas de reserva representan el principal componente proteico del endospermo. Son divididas en dos grupos diferenciados: gluteninas y gliadinas, las cuales constituyen el núcleo central del gluten, malla proteica responsable de las propiedades viscoelásticas de la masa de harina de trigo (Fig. 3).

Figura 3. Después del amasado la masa de harina puede ser estirada demostrando sus propiedades viscoelásticas (Shewry 2009).

Shewry et al. (1986) las clasificaron en tres principales grupos en función de sus características estructurales:

- Prolaminas de alto peso molecular: son las subunidades de alto peso molecular de glutenina.
- Prolaminas pobres en azufre: corresponden a las ω-gliadinas.

- Prolaminas ricas en azufre: α-, β-, γ-gliadinas y subunidades de bajo peso molecular de glutenina.

En base a la movilidad en un sistema de electroforesis a pH ácido (A-PAGE), las gliadinas se separan en: α-, β-, γ- y ω-gliadinas. Estas proteínas son codificadas por genes localizados en el brazo corto de los cromosomas homeólogos 1 y 6. El loci *Gli-1* del grupo homeólogo 1 (*Gli-A1*, *Gli-B1* y *Gli-D1*) controla la mayoría de las γ- y ω-gliadinas, mientras que el loci *Gli-2* (*Gli-A2*, *Gli-B2* y *Gli-D2*) del grupo homeólogo 6 controla todas las α/β-gliadinas (Payne et al. 1982). Cada uno de los loci está formado por varios genes estrechamente ligados.

Las gliadinas tienen un tamaño comprendido entre 30 y 75 kDa. Son proteínas monoméricas y presentan uniones tipo puentes de hidrógenos e interacciones hidrofóbicas. Las α/β-gliadinas y las γ-gliadinas presentan 6 y 8 residuos de cisteína, respectivamente, que forman puentes disulfuro intramoleculares, mientras que las ω-gliadinas no contienen ningún residuo de cisteína. Una pequeña proporción de α/β y γ-gliadinas puede presentar un número impar de cisteínas debido a alguna mutación que les permite incorporarse al gluten mediante uniones intermoleculares y actuar como terminadores de la polimerización (Barak et al. 2015). Estas proteínas han sido estudiadas, no sólo por su importancia en las propiedades viscoelásticas del gluten, al ser responsables de la cohesividad y extensibilidad de la masa, sino también por su elevado nivel de polimosfismo que las hace un interesante marcador de variación genética (Lafiandra et al. 1990), pudiendo usarse para la identificación de cultivares en trigo (Bushuk y Zillman 1978).

Las gluteninas, por su parte, son responsables de la elasticidad o tenacidad de la masa y, de acuerdo a su separación en electroforesis desnaturalizante a pH básico (SDS-PAGE), se dividen en subunidades de alto peso molecular (HMWGs) y subunidades de bajo peso molecular (LMWGs) (Payne 1987). A diferencia de las gliadinas son proteínas poliméricas, encontrándose las subunidades unidas covalentemente mediante puentes disulfuro (S-S).

Las HMWGs están codificadas por el loci *Glu-1* localizado en el brazo largo de cada uno de los cromosomas del grupo homeólogo 1, denominándose *Glu-A1*, *Glu-B1* y *Glu-D1*, respectivamente (Payne et al. 1980). Cada locus consiste en dos genes estrechamente ligados que codifican dos subunidades (*x* e *y*) que difieren en su peso molecular (80-100 kDa) y, por tanto, presentan diferente movilidad en SDS-PAGE. La subunidad de tipo *x*, muestra mayor peso molecular y menor movilidad en SDS-PAGE,

mientras que la subunidad tipo *y*, es más pequeña y presenta mayor movilidad electroforética (Harberd et al. 1986). Ambas subunidades presentan la misma estructura: péptido señal, dominio N-terminal, dominio repetitivo y dominio C-terminal (Fig. 4). La mayoría de las subunidades *x* poseen cuatro residuos de cisteína, mientras que la mayoría de tipo *y* poseen siete residuos de cisteína. Dos residuos de las subunidades *x* forman un puente disulfuro intramolecular, mientras que los otros dos formarían puentes disulfuro intermoleculares con subunidades de tipo *y*. Hasta el momento, para las subunidades de tipo *y* se han descrito dos puentes disulfuro intermoleculares paralelos entre dos diferentes subunidades *y* y otro que las une a las LMWGs (Shewry et al. 2002).

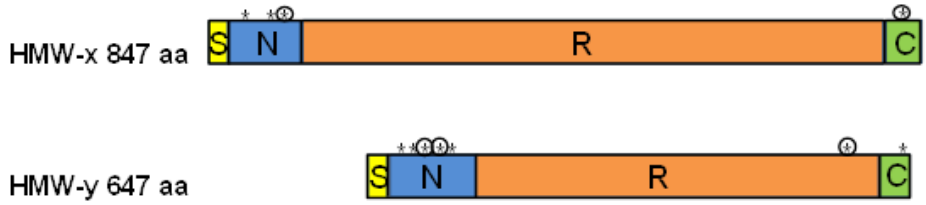


Figura 4. Esquema representativo de las HMWGs. S: péptido señal; N: dominio N-terminal; R: dominio repetitivo; C: dominio C-terminal. Los residuos de cisteína están indicados con asteriscos. Los asteriscos resaltados representan residuos de cisteína involucrados en puentes disulfuro intermoleculares.

Por otro lado, las LMWGs presentan pesos moleculares entre 30-50 kDa y están codificadas por una familia multigénica localizada en el loci *Glu-3* del brazo corto de los cromosomas 1A, 1B y 1D (Singh y Shepherd 1988; Pogna et al. 1990), el cual está estrechamente ligado con los loci *Gli-1* que codifican las γ- y ω- gliadinas. En función de su movilidad en SDS-PAGE y su punto isoeléctrico, Jackson et al. (1983) las dividieron en B-, C- y D-LMWGs. De todas ellas, las B-LMWGs son las que presentan el mayor número de subunidades y han sido asociadas a la calidad del trigo duro, utilizado para la elaboración de la pasta (Wrigley et al. 2006). Estas subunidades están clasificadas dentro de tres grupos en función del primer aminoácido presente en la proteína madura: LMW-i (isoleucina), LMW-m (metionina) y LMW-s (serina) (D'Ovidio y Masci 2004). Los genes para las subunidades LMW-m y LMW-s han sido principalmente identificados en los loci *Glu-B3* y *Glu-D3* y los genes para las LMW-i en el locus *Glu-A3*. Sin embargo, algunos estudios han mostrado la presencia de genes de LMW-m en *Glu-A3* (Lee et al. 1999a; Dong et al. 2010; Zhang et al. 2013). Asimismo, en un reciente estudio se identificó un pseudogen de LMW-s en el locus *Glu-A3* (Luo et al. 2015). La estructura de estas subunidades consiste de cuatro dominios: péptido señal, dominio N-terminal,

dominio repetitivo y dominio C-terminal (Fig. 5). Este último formado por tres subdominios (C-I, C-II y C-III). Presentan 6+2 residuos de cisteína que forman tres puentes disulfuro intramoleculares y dos intermoleculares. Sin embargo, las subunidades LMW-i poseen una estructura única; al carecer del dominio N-terminal, el dominio repetitivo es excluido en la formación de los puentes disulfuro intermoleculares en el gluten (D'Ovidio y Masci 2004).

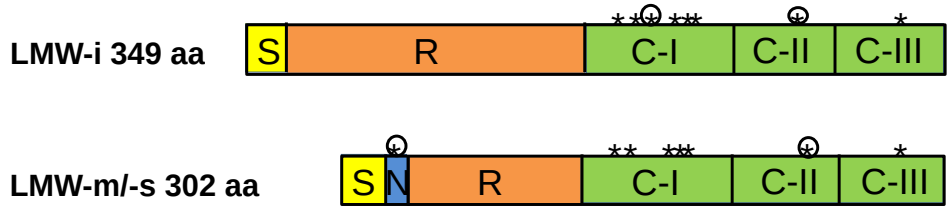


Figure 5. Esquema representativo de las LMWGs. S: péptido señal; N: dominio N-terminal; R: dominio repetitivo; C-I, C-II y C-III: dominios C-terminal. Los residuos de cisteína están indicados con asteriscos. Los asteriscos resaltados representan residuos de cisteína involucrados en puentes disulfuro intermoleculares.

En un principio, debido a la complejidad de esta familia multigénica, su caracterización había sido enfocada en menor medida que otras prolaminas. Sin embargo, recientes estudios han aportado nuevos datos para esclarecer su composición, como los de Zhang et al. (2013) que identificaron más de 15 genes en entradas individuales de trigo harinero agrupados en diferentes grupos: dos grupos (m_{AD} y i_A) para el locus *Glu-A3*, otros dos (m_{BD} y s_{BD}) para el locus *Glu-B3* y otro grupo (m_D-2-s_{BD}) junto con algunos genes no vinculados (m_{BD}, m_{AD} y m_D-1) para el locus *Glu-D3*. No obstante, el número exacto de genes sigue siendo desconocido (Rasheed et al. 2014).

El gluten, además de ser responsable de las propiedades viscoelásticas de la masa, haciéndola adecuada para la elaboración de diversos productos, es responsable de la enfermedad celíaca, una condición inflamatoria mediada por las células T que causa daños en el revestimiento del intestino delgado. Los principales epítopos reactivos han sido identificados en las gliadinas, principalmente en las α-gliadinas (Rasheed et al. 2014). Sin embargo, algunos estudios también han asociado a las gluteninas con esta enfermedad debido a la presencia de epítopos reactivos como los encontrados en las gliadinas (Vader et al. 2002).

Las mutaciones que afectan a la estructura de las proteínas de reserva pueden dar lugar a diferencias funcionales. Para las LMWGs, subunidades con residuos extra de

cisteína han mostrado un incremento en la fuerza de la masa por aumento del tamaño del polímero de gluten, incrementando de esta manera la elasticidad de las mismas (Xu et al. 2006). Además, los dominios repetitivos largos proporcionan un alto contenido de glutaminas disponibles para interacciones intermoleculares a través de puentes de hidrógeno que podrían fortalecer las masas (Masci et al. 1998, 2000). Por lo tanto, es importante la búsqueda de nuevas variantes de prolaminas con el objetivo de aumentar el fondo genético del trigo, ya que suponen una buena herramienta para la mejora de la calidad de la harina o sémola de trigo por su importante contribución en las propiedades viscoelásticas de la masa. Sin embargo, su papel en la enfermedad celíaca hace necesario la selección de materiales menos reactivos con el fin de conseguir variedades aptas para celíacos. Por otra parte, su elevado polimorfismo las hace también buenas candidatas como marcador de análisis de la diversidad genética.

Proteínas sintetizadoras del almidón

El almidón constituye aproximadamente el 70% del peso seco del grano de trigo y tiene un importante efecto sobre la calidad de los productos alimentarios, ya que afecta directamente a su apariencia, estructura y otras características. Está formado por dos componentes: amilosa y amilopectina. La amilosa es una molécula lineal formada por residuos D-glucosa unidos por enlace α-1,4, mientras que la amilopectina es una molécula de cadena más larga y ramificada, formada por cadenas de D-glucosa unidas por enlace α-1,4 y ramificaciones α-1,6 cada 10-15 residuos. En trigo, normalmente, el almidón contiene un 20-30% de amilosa y 70-80% de amilopectina. Las propiedades físicas y químicas del almidón (gelidificación, pegajosidad y gelatinización), y consecuentemente la calidad del producto final dependen de la proporción de amilosa y amilopectina (Zeng et al. 1997).

En el endospermo, la síntesis del almidón se produce dentro del amiloplasto a partir de glucosa-1-fosfato. La primera enzima en actuar es la ADP glucosa pirofosforilasa que sintetiza adenosina difosfato glucosa (ADPG), el sustrato para el resto de enzimas. Entre las principales almidón-sintasas que actúan posteriormente, se encuentran: la almidón-sintasa unida al gránulo I (GBSSI o waxy; 59 o 61 kDa), la almidón sintasa I (SSI o SGP-3; 80 kDa), la almidón-sintasa II (SSII o SGP-1; 100-115 kDa) y la enzima ramificadora del almidón I (SBEI o SGP-2; 92 kDa). La proteína waxy es la única responsable de la síntesis de amilosa, mientras que el resto de sintasas están relacionadas con la síntesis de amilopectina. Yamamori and Endo (1996) indicaron que

los genes para las proteínas waxy, SGP-1 y SGP-3 están localizados en el brazo corto de los cromosomas 7A, 7B y 7D, con la excepción del gen *waxy* del genoma B que está localizado en el cromosoma 4AL debido a una translocación de parte del cromosoma 7BS (Chao et al. 1989).

De todas las enzimas implicadas en la síntesis del almidón las waxy han sido las más estudiadas. Estas proteínas son muy importantes en la industria alimentaria asiática, ya que están asociadas a la fabricación de diversos tipos de fideos conocidos como "noodles", los cuales requieren porcentajes de amilosa más bajos de lo normal para obtener almidones con mayor hinchamiento y menor firmeza que los de la pasta (Peña 2002). Diversos estudios han mostrado que el nivel de variabilidad de estas proteínas es bajo, más incluso si las comparamos con otras proteínas del grano como las proteínas de reserva. No obstante, se han detectado algunos alelos con diferente movilidad en SDS-PAGE y con diferente punto isoeléctrico e incluso alelos nulos (Guzmán y Alvarez 2015). Éstos últimos son de gran importancia ya que permiten la creación de los trigos carentes de amilosa o trigos *waxy* (Nakamura et al. 1995). También se ha estudiado la importancia relativa de cada una de las subunidades waxy, demostrándose que la proteína Wx-B1 es la que juega el papel más importante, así como el efecto de algunas mutaciones que reducen la actividad enzimática, como el alelo *Wx-D1f* (Yanagisawa et al. 2001).

En los últimos años el estudio de la calidad del almidón ha cobrado importancia en la industria occidental panadera, ya que se ha comprobado que afecta a procesos como el enranciamiento, absorción del agua y los tiempos de fermentación de las masas, e incluso la digestibilidad de los almidones. Además, la calidad del almidón puede ser importante fuera de la industria alimentaria, como en la generación de bioetanol a partir del grano, donde se ha demostrado un mayor rendimiento utilizando trigos *waxy* (Wu et al. 2006).

Recursos fitogenéticos

Con el comienzo de la Agricultura se desarrollaron fuertes presiones selectivas sobre las plantas cultivadas que conllevaron una progresiva diferenciación de las formas silvestres de procedencia, lo cual culminó en la domesticación de muchas de estas plantas y su adaptación a nuevos ambientes, manejos y usos (Cubero 2003). Durante este proceso, el hombre seleccionó aquellas semillas que mejor se adaptaban a los pretendidos usos y, consecuentemente, sólo una parte de la diversidad genética presente en las formas silvestres pasó a las formas cultivadas.

La diversidad genética de las plantas cultivadas fue mantenida durante un largo periodo de tiempo, pero en los últimos años la agricultura moderna ha dado lugar a una mayor uniformidad genética de los cultivos (Esquinas-Alcázar 2005). Con la Revolución Verde, el cultivo de unas pocas variedades de alto rendimiento y adaptadas a las condiciones de agricultura intensiva fue priorizada respecto a la producción diversificada y las variedades locales o tradicionales fueron siendo relegadas. Por consiguiente, se ha producido un estrechamiento de la base genética de los cultivos. La reducción de la variabilidad genética ha incrementado la vulnerabilidad de los cultivos debido a la pérdida de genes de interés necesarios para enfrentarse a nuevas plagas y enfermedades o condiciones ambientales adversas (Esquinas-Alcázar 2005).

Esta variabilidad genética es la base de la selección natural y el *pool* genético utilizado por los mejoradores para el desarrollo de nuevas variedades. La continua pérdida de variabilidad ha promovido la adopción de diversas estrategias a fin de conservar la diversidad genética (Esquinas-Alcázar 2005). Existen dos sistemas de conservación: *ex situ* e *in situ*, destinados a la conservación de variedades locales y especies silvestres relacionadas con los cultivos (Hammer et al. 2001). El principal sistema es la conservación *ex situ* en Bancos de germoplasma a través de la conservación en forma de semillas, plantas vivas, tejidos vegetales cultivados *in vitro*, polen y ADN. La conservación *in situ* involucra la protección de los ecosistemas donde las plantas de interés han desarrollado sus características. No obstante, si bien este tipo de conservación ha sido mayoritariamente asociada a plantas de escaso o nulo interés agrícola, a partir del desarrollo del Tratado Internacional sobre los Recursos Fitogenéticos para la Alimentación y la Agricultura (FAO 2001), se ha puesto en valor una variante de este tipo de conservación *in-situ* denominada conservación *on-farm* o en cultivo asociada al mantenimiento y conservación de variedades locales mediante técnicas tradicionales por parte de los propios agricultores. Ambos métodos son complementarios y el uso simultáneo de ambos podría asegurar la correcta conservación de los recursos genéticos (Hammer 2003).

Otro aspecto importante es la caracterización y evaluación de los recursos fitogenéticos para su apropiada conservación y uso (Rao y Hodgkin 2002). Es necesario un cribado previo para la correcta elección de los materiales a conservar. Además, numerosas entradas están almacenadas en los Bancos de Germoplasma y la mayoría son infrautilizadas debido a la inexistencia de programas eficaces que las evalúen para caracteres de interés (Esquinas-Alcázar 2005). Por lo tanto, son necesarios proyectos que

pongan a punto el valor de los recursos fitogenéticos para que puedan ser utilizados en programas de mejora y satisfacer las demandas actuales.

Especies silvestres relacionadas y trigos abandonados como recursos fitogenéticos

En 1971, Harlan y deWet desarrollan un sistema de clasificación que define tres niveles de interrelación entre las plantas en función de la posibilidad de transferir información genética entre ellos. Estos niveles, denominados como «*gene pools*», se establecen a nivel primario (GP-1), secundario (GP-2) y terciario (GP-3). El GP-1 estaría constituido por las formas que pueden cruzarse fácilmente entre sí, que en el caso del trigo estaría constituido por todas aquellas formas primitivas y variedades locales de trigo, cuya cruzabilidad con los cultivos de trigo moderno no presentan dificultad; el GP-2 incluye aquellas entre las cuales el cruzamiento es posible, pero los híbridos pueden presentar ciertas dificultades, solucionables con algún esfuerzo. En el caso del GP-3, las dificultades son mucho mayores y la transferencia necesita de técnicas especiales.

Si bien generalmente los materiales utilizados en la mejora del trigo moderno han comprendido mayoritariamente a materiales incluidos dentro del GP-1 como trigos en desuso o materiales ya mejorados dentro de la especie; en los últimos años, la búsqueda de nueva variabilidad genética utilizable para incrementar su base genética se ha extendido a los parientes del trigo moderno, incluidos en el GP-2, cuyos genomas o bien están relacionados o bien se han identificado como donantes de los genomas presentes en el trigo. Estos parientes comprenden diversas especies incluidas dentro de los géneros *Triticum* y *Aegilops*. Se da la paradoja que durante algunos años, estas especies del género *Aegilops* fueron clasificadas taxonómicamente como pertenecientes al género *Triticum* (Kimber y Sears 1987).

Las especies del género *Aegilops* se incluyen mayoritariamente dentro del GP-2 del trigo, con excepción de *Ae. tauschii* que al ser el donador del genoma D de trigo presenta una gran homología con éste y es englobado en el GP-1 (Dvorak et al. 1998). Estas especies se distribuyen de forma natural por áreas circunmediterráneas y el Próximo Oriente, detectándose la mayor diversidad en zonas del Creciente Fértil (Kole 2011). Las relaciones de los genomas de las especies de *Aegilops* con los genomas del trigo son de gran interés debido al potencial de estas especies silvestres como fuente de nueva variabilidad que podría ser transferida a trigo. Wang et al. (2011a) analizando el locus *Glu-3*, sugirieron que las especies con genoma C y U podrían estar relacionadas con el genoma A del trigo, mientras que aquellas con genoma M y N lo estarían con el

genoma D del trigo. Sin embargo, Manson-Gamer et al. (1998) estudiando los genes *Wx*, indicarón que todos estos genomas (C, M, N y U) estarían relacionados con el genoma D. Por su parte, los genomas S de las especies de *Aegilops* de la sección *Sitopsis* estarían relacionados con el genoma B del trigo (Haider et al. 2013). Dentro de esta sección se encuentra *Ae. speltoides*, el principal candidato propuesto como donador del genoma B (Petersen et al. 2006).

Aunque el uso de especies silvetres ha sido denostado por algunos mejoradores debido a la incorporación no deseada de numerosos caracteres indeseables (*linkage drag*), no es menos cierto que dado el potencial que estas especies han mostrado como fuentes de genes relacionados con caracteres de interés como resistencia a enfermedades, tolerancia al estrés y calidad (Schneider et al. 2008), estos materiales suponen una oportunidad para la mejora del trigo de cara a su diversificación y adaptación a nuevos usos y ambientes.

Todas las especies del género *Triticum* forman parte del GP-1, entre las cuales se encuentran las especies diploides: *T. urartu* y escaña (silvestre y cultivada), que están relacionadas con el genoma A de los trigos poliploides y, por tanto, podrían ser una fuente de variabilidad muy importante para la mejora del trigo. Morfológicamente los tres trigos son muy parecidos, aunque *T. urartu* está aislado reproductivamente de la escaña (Johnson y Dhaliwal 1976; Sharma y Waines 1981) y su distribución natural está prácticamente confinada en el Próximo Oriente.

La escaña cultivada forma parte de los trigos vestidos, que están caracterizados por poseer glumas fuertemente adheridas al grano que pueden incluso permanecer después de la trilla. Fue un trigo muy importante en la Antigüedad, siendo uno de los cultivos que se consideran asociados al nacimiento de la Agricultura en el Creciente Fértil (Zohary y Hopf 1988). Si bien fue extensamente cultivado en zonas del Oriente Medio, Asia central, Europa y Norte de África; hoy día se considera un cultivo abandonado, puesto que ha sido reemplazado por variedades modernas de alto rendimiento, cultivándose sólo en algunas zonas aisladas de Turquía, Cáucaso, Europa y Marruecos (Zaharieva y Monneveux 2014). No obstante, actualmente se ha producido un resurgir de los cultivos de trigos antiguos infrautilizados o abandonados, entre los que además de la escaña se encuentran el emmer cultivado y el espelta; estos se han asociado a una agricultura de tipo más tradicional, incluido el cultivo ecológico, valorándose sus posibles beneficios saludables y/o nutricionales. Paralelamente, desde el punto de vista de

la mejora de trigo moderno, estos trigos son una importante fuente de nueva variabilidad para genes de interés (Zaharieva y Monneveux 2014).

En los últimos años la búsqueda de variación en especies de formas primitivas de trigo, como son *T. urartu* y la escaña tanto en su forma silvestre como cultivada, y especies silvestres relacionadas, como las especies de *Aegilops*, ha sido una parte importante del trabajo de nuestro grupo. Nueva variabilidad ha sido encontrada para proteínas relacionadas con la calidad que podría ser útil en programas de mejora de trigo. La variabilidad en proteínas de reserva ha sido extensamente analizada, fundamentalmente a nivel de proteína, tanto en escaña cultivada como en *T. urartu* (Alvarez et al. 2006, 2013; Caballero et al. 2008, 2009; Martín et al. 2008). Respecto a los genes waxy, se ha detectado una gran variación genética en especies de *Aegilops*, escaña cultivada y *T. urartu* (Ortega et al. 2014a,b). Por último, los menos estudiados han sido los genes *Pin*, con sólo algunas entradas de *T. urartu* y escaña caracterizadas y secuenciadas (Guzmán et al. 2011).

Estos trabajos han mostrado que la caracterización de estas especies es crucial para la búsqueda de nueva variación que podría ser usada para la mejora del trigo moderno. Sin embargo, un análisis más profundo es necesario en la caracterización de las proteínas de reserva, añadiendo el análisis de su caracterización genética. Para las LMWGs, esta caracterización podría ser ampliada tanto en especies de *Aegilops* como en especies de escaña y *T. urartu*. Respecto a los genes *Pin* la secuenciación en especies de *Aegilops* podría revelar nueva y útil variación.

Objetivo

El objetivo general de esta Tesis Doctoral ha sido la evaluación y caracterización de la variación alélica detectada en genes implicados en la dureza del grano, *Pin* y *Gsp-1*, así como en los genes de LMWGs en especies diploides de los géneros *Aegilops* y *Triticum*, de cara a su valoración como posibles recursos en la mejora genética del trigo moderno.

CAPÍTULO I

DIVERSIDAD ALÉLICA Y MOLECULAR DE GENES DE *PUROINDOLINAS* EN CINCO ESPECIES DIPLOIDES DEL GÉNERO *Aegilops*

Publicado como:

S. Cuesta, C. Guzmán, J.B. Alvarez (2013) Allelic diversity and molecular characterization of *Puroindoline* genes in five diploid species of the *Aegilops* genus. *Journal of Experimental Botany* **64**: 5133-5143.

Resumen

La dureza del grano es una importante característica de la calidad del trigo. Esta característica está relacionada con la variación, y la presencia, de las puroindolinas (PINA y PINB). Esta variación puede ser incrementada por el polimorfismo alélico presente en las especies de *Aegilops* que están relacionadas con el trigo. Este estudio evaluó la variabilidad alélica en los genes *Pina* y *Pinb* en cinco especies diploides del género *Aegilops*, junto con la caracterización molecular de las principales variantes alélicas encontradas en cada especie. Este polimorfismo resultó en 16 alelos para el gen *Pina* y 24 alelos para el gen *Pinb*, de los cuales 10 y 17, respectivamente, fueron nuevos. Diversas mutaciones fueron detectadas en las proteínas maduras de estos alelos, las cuales podrían influir en las características de dureza de estas proteínas. Este estudio mostró que las especies diploides del género *Aegilops* podrían ser una buena fuente de variabilidad genética para los genes *Pina* y *Pinb*, las cuales podrían ser usadas en programas de mejora para extender el rango de diferentes texturas en trigo.

Palabras clave: *Aegilops* sp., diversidad, dureza del grano, no AUG como codón de inicio, puroindolinas, trigo.

Abstract

Grain hardness is an important quality trait in wheat. This trait is related to the variation in, and the presence of, puroindolines (PINA and PINB). This variation can be increased by the allelic polymorphism present in the *Aegilops* species that are related to wheat. This study evaluated allelic *Pina* and *Pinb* gene variability in five diploid species of the *Aegilops* genus, along with the molecular characterization of the main allelic variants found in each species. This polymorphism resulted in 16 alleles for the *Pina* gene and 24 alleles for the *Pinb* gene, of which 10 and 17, respectively, were novel. Diverse mutations were detected in the deduced mature proteins of these alleles, which could influence the hardness characteristics of these proteins. This study shows that the diploid species of the *Aegilops* genus could be a good source of genetic variability for both *Pina* and *Pinb* genes, which could be used in breeding programmes to extend the range of different textures in wheat.

Keywords: *Aegilops* sp., diversity, grain hardness, puroindolines, non-AUG start codon, wheat.

Introduction

Hardness or grain endosperm texture is an important quality trait in bread wheat (*Triticum aestivum* L. ssp. *aestivum;* 2n = 6× = 42, AABBDD). This characteristic depends on the level of adhesion between the protein matrix and the starch granules in the endosperm. Bread wheat is classified as being either soft or hard according to this trait and this trait, therefore, determines to what use the bread wheat is put (Morris and Rose 1996).

Several studies have shown a direct relationship between hardness and puroindoline type and content (Morris 2002). Puroindolines (PINA and PINB) are two basic grain proteins that have a mol. wt of ~13 kDa, a conserved cysteine backbone formed by ten cysteine residues, and a tryptophan-rich hydrophobic domain (Blochet et al. 1993; Gautier et al. 1994). The genes encoding these proteins in bread wheat (*Pina-D1* and *Pinb-D1*) are located at the locus *Ha* (*Hardness*), which is on the short arm of chromosome 5D (Morris 2002). Ten genes, in an 82,353 bp region, have been described at this locus, including *GSP-1* (*Grain softness protein)*. However, to date, it has only been possible to establish the role of two of them (*Pina-D1* and *Pinb-D1*) on grain texture (Chantret et al. 2005; Li et al. 2008a).

In wheat, puroindolines have been detected in soft and hard bread wheat, but are not present in durum wheat (*T. turgidum* ssp. *durum* Desf. em. Husn.; 2n = 4× = 28, AABB) and other tetraploid species, which are very hard wheats (Gautier et al. 2000). In the latter cases, several studies have shown that both *Pin* genes are absent because they were deleted from chromosomes 5A and 5B during the polyploidization event (Chantret et al. 2005). However, both genes are present in the diploid species that have been identified as putative donors of the A and B genome in polyploid wheats (Gautier et al. 2000; Massa et al. 2006). These are *T. urartu* Thum. ex Gandil. (2n = 2× = 14, AA) for the A genome and *Ae. speltoides* Tausch (2n=2×=14, SS) for the B genome (Petersen et al. 2006).

In bread wheat, these genes have been recovered from the D genome of *Ae. tauschii* Coss. (Li et al. 2008a). The presence of wild puroindoline alleles (*Pina-D1a, Pinb-D1a)* produces the soft phenotype, whereas mutations that lead to amino acid changes or loss of protein, result in a hard phenotype (Giroux and Morris 1997, 1998;

Lillemo and Morris 2000; Morris 2002). This has driven the search for novel puroindolines alleles over the last few years, which could result in grain hardness changes and an increase in the range of available textures (for a review, see Bhave and Morris 2008a). More recently, this search has been extended to relatives of wheat, including species of the *Aegilops* genus (Gautier et al. 2000; Darlington et al. 2001; Lillemo et al. 2002; Massa et al. 2004; Chen et al. 2005; Gazza et al. 2006; Simeone et al. 2006; Guzmán et al. 2011).

The *Aegilops* genus consists of 22 species distributed in five sections: Sect. *Aegilops* L.; Sect. *Comopyrum* (Jaub. & Spach) Zhuk.; Sect. *Cylindropyrum* (Jaub. & Spach) Zhuk.; Sect. *Sitopsis* (Jaub. & Spach) Zhuk., and Sect. *Vertebrata* Zhuk. emend. Kihara (van Slageren 1994). Ten of these species are diploid (2n = 2× = 14 chromosomes) with up to 10 different genomes: *Ae. bicornis* (Forssk.) Jaub. & Spach (S^b genome), *Ae. comosa* Sm. (M genome), *Ae. longissima* Schweinf. & Muschl. (S^l genome), *Ae. markgrafii* (Greuter) K. Hammer (C genome), *Ae. searsii* Feldman & Kislev ex K. Hammer (S^s genome), *Ae. sharonensis* Eig (S^{sh} genome), *Ae. speltoides* Tausch (S genome), *Ae. tauschii* (D genome), *Ae. umbellulata* Zhuk (U genome), and *Ae. uniaristata* Vis. (N genome). The rest of the species show combinations of these genomes. The relationships between these genomes have also been evaluated, although some analyses are clearly contradictory to each other. All evidence suggests that the D genome of *Ae. tauschii* is related to the D genome of bread wheat, and the S (S, S^b, S^l, S^s and S^{sh}) genomes are related with the B genome to a greater or lesser degree. However, the phylogenetic relationships between the rest of the genomes are not clear. Wang et al (2011a) using the *Glu-3* loci suggested that the C and U genomes could be related to the A genome, while the M and N genomes could be related to the D genome. In contrast, Mason-Gamer et al (1998), using the *Wx* genes sequences, indicated that all these genomes (C, M, N and U) were related to the D genome.

Consequently, because bread wheat contains the *Pin* genes from the D genome, the allelic variation in species with genomes different to the D genome could be useful when attempting to diversify wheat grain texture through the introgression of new alleles of both *Pin* genes. Previous studies have highlighted the potential of these genomes to change the texture of durum wheat. The newly discovered alleles can be transferred to cultivated wheats, as was shown by Gedye et al. (2004), who evaluated a population of 75 CIMMYT synthetic hexaploid wheats with *Ae. tauschii* genotypes carrying different *Pin* alleles. This resulted in wheats with softer texture than common ones. A similar

effect has been observed in tritordeum (×*Tritordeum* Ascherson and Graebner, amphiploid between durum wheat and *Hordeum chilense* Roem. et Schult.), which has a grain hardness similar to that of bread wheat (Alvarez et al. 1992). This change in the grain texture could be related to the presence of the *Hin* genes, orthologous of *Pin* genes, from *H. chilense* (Yanaka et al. 2011; Terasawa et al. 2012). For this reason, the characterization of related genes in wheat relatives could be very important, since the variability detected could be used to increase the genetic and phenotypic variability of this trait in common wheat.

The aim of the current study was to evaluate the allelic variability of *Pina* and *Pinb* genes of five diploids species from the *Aegilops* genus and to characterize molecularly the main allelic variants found in each species.

Materials and Methods

Plant material

Eighty-four accessions of *Aegilop*s were used in this study, obtained from the National Small Grains Collection (Aberdeen, Idaho, USA). These accessions belong to five diploid species of the *Aegilops* genus: *Ae. comosa*, *Ae. markgrafii*, *Ae. searsii*, *Ae. speltoides* and *Ae. umbellulata*, where at least 15 accessions of each species were analysed (Supplementary Table 1).

DNA isolation and PCR amplification of Pina and Pinb

Genomic DNA was isolated from young leaves of a single plant per accession according to the method to the cetyltrimethyl ammonium bromide (CTAB) method (Stacey and Isaac 1994). Full-length *Pina* and *Pinb* genes were amplified with gene-specific primers (Table 1). Reactions were performed in 15 μl of total volume containing 50 ng of genomic DNA, 0.4 μM of each primer, 0.2 mM of dNTPs, 1.5 mM of $MgCl_2$, 1× of reaction buffer and 0.75 U of *Taq* DNA polymerase (Promega, Madison, WI, USA). The amplification was performed with an initial denaturation step of 3 min at 94 °C followed by 35 cycles as follows: 45 s at 94 °C, a step of annealing between 30 s and 90 s at 54 °C or 58 °C for *Pina*, and between 30 s or 60 s at 54 °C or 64 °C for *Pinb*, then 45 s at 72 °C. To finish, a final extension step at 72 °C for 5 min was performed. The PCR products were separated by electrophoresis on polyacrylamide gels of 8% (p/v; C: 1.28%), stained with ethidium bromide and visualized under UV light.

Table 1. Gene-specific PCR primers to produce full-length sequences of the genes *Pina* and *Pinb*.

Gene	Primers Sequence (5′ → 3′)*	Amplified species
Pina	MS-F: GGTGTGGCCTCATCTCATCT	*Ae. comosa*
	MS-R: AAATGGAAGCTACATCACCAGT	*Ae. speltoides*
		Ae. searsii
		Ae. umbellulata
Pina	CG-F: CCACCTGCACCAAACACATGA	*Ae. markgrafii*
	CG-R: CCACCTGCACCAAACACATGA	
Pinb	MS-F: AATAAAGGGGAGCCTCAACC	*Ae. comosa*
	MS-R: CGAATAGAGGCTATATCATCACCA	*Ae. markgrafii*
		Ae. searsii
		Ae. speltoides
		Ae. umbellulata

* MS primers from Massa et al. (2004); and CG primers designed by us for this study.

Cloning and sequencing

PCR products were purified and ligated into pGEM-T easy vector (Promega) and cloned in *Escherichia coli* JM109 competent cells. Three positive clones for each PCR product were sequenced in both directions with M13 universal primers using an ABI Prism 310 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The sequences obtained were analyzed and compared using the Geneious Pro ver. 5.0.3 software (Biomatters Ltd.). The novel sequences are available from Genbank database.

Data analysis

DNA analyses were conducted by DNAsp ver. 5.0 (Librado and Rozas 2009) and parameters as total number of mutations (η), average number of nucleotide differences (k), and number of polymorphic sites (s) were calculated. DNA sequences of allelic wild-type *Pina-D1a* and *Pinb-D1a* of bread wheat cv. Chinese Spring (CS) described by Gautier et al. (1994), were included for comparison. Nucleotide diversity was estimated as theta (θ), the number of segregating (polymorphic) sites (Watterson 1975), and pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987). Tests of neutrality were performed using Tajima's D statistic (1989).

Results

Amplification and sequencing of puroindoline genes

The allelic variability of the *Pina* and *Pinb* genes in the *Aegilops* accessions studied are shown in Figs 1 and 2, respectively. The polymorphism observed was large

for both genes, although it was higher for *Pinb* than for *Pina*. With regards to the *Pina* gene, all accessions produced amplicons and no heterozygous accession was detected (the *Gsp-1* gene was amplified simultaneously due to its high homology with *Pina,* although it was not the subject of the study). In contrast, for *Pinb*, some accessions were classified as null because they did not show any amplicons for this gene. This was detected in three accessions of *Ae. comosa* and *Ae. umbellulata*. Additionally, four heterozygous accessions for *Pinb* were detected (one each for *Ae. comosa* and *Ae. speltoides* and two for *Ae. searsii*). Therefore, the total number of puroindoline alleles in each species ranged from two to four for *Pina* and three to nine for *Pinb*. The alleles were classified according to their gene and genome type and were assigned a Roman numeral.

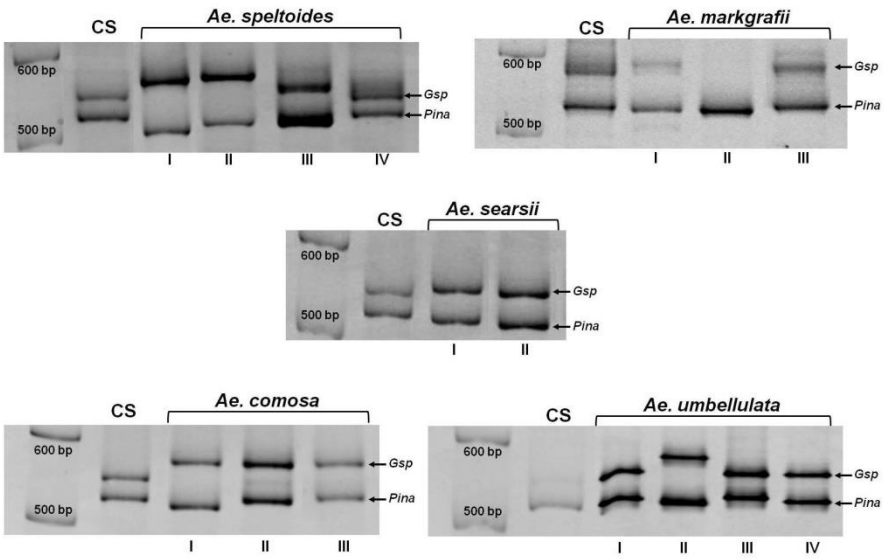


Figure 1. Allelic variation for the *Pina* gene detected in the *Aegilops* species evaluated. Bread wheat cultivar Chinese Spring (CS) was used as control.

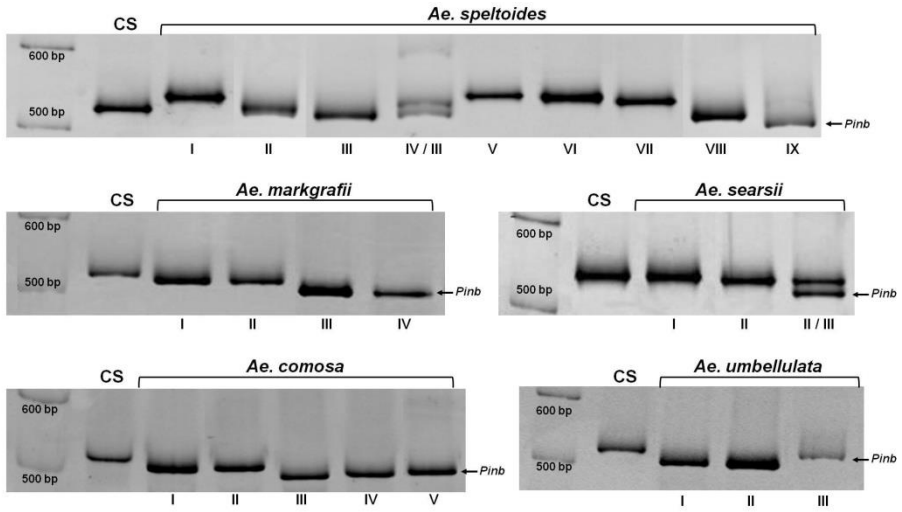


Figure 2. Allelic variation for the *Pinb* gene detected in the *Aegilops* species evaluated. Bread wheat cultivar Chinese Spring (CS) was used as control.

Each allele that was identified by electrophoretic mobility was further confirmed by sequencing. For some alleles, more than one accession was selected for sequencing analysis in order to confirm the electrophoretic results or to detect further variability that was not detected by electrophoretic analysis. The sequencing analysis showed that some alleles, which seemed to be equal in terms of electrophoresis mobility, were in fact different. This occurred in two accessions of *Ae. umbellulata* for *Pina*. In total, 53 amplicons were isolated and sequenced (Tables 2 and 3). Of these, 16 different alleles were found for the *Pina* gene and 24 for the *Pinb* gene. Of the alleles identified and sequenced in this study, 27 (10 for *Pina* and 17 for *Pinb*) were identified as being novel after comparison with other genes available in GenBank (Tables 2 and 3).

The different coding sequences obtained were aligned with the wild-type alleles: *Pina-D1a* and *Pinb-D1a* found in the bread wheat cv. Chinese Spring. In the alignments of the overall *Pina* coding sequences (447 bp), no insertions or deletions were observed among any of the *Aegilops* species. However, the *Pina-*S^s*1-II* allele of *Ae. searsii* possessed a shorter coding region (444 bp) due to a point mutation that resulted in a change in the final codon (TGG) to a stop codon (TAG). Furthermore, some of the new alleles showed similar nucleotide sequences to other alleles previously described in different species. The *Pina-U1-I* allele of *Ae. umbellulata* was identical to a sequence that had been previously described in *Ae. columnaris* (NCBI: AY608598; Huo et al.

unpublished data); the *Pina-M1-II* allele had the same sequence as *Pina-C1-III* allele and *Pina-S1-IV* was the same as *Pina-U1-III*.

Table 2. *Pina* alleles in diploids species of the *Aegilops* genus.

Species	Allele	Size		Accession (PI)	NCBI accession	Previous NCBI accessions [references]
		DNA (bp)	Protein (aa)			
Ae. comosa	*Pina-M1-I*	447	148	551018	JX648367	FJ898209, FJ898210, FJ898211 [1]
				542176	FJ898209	
				551022	JX648368	
	Pina-M1-II	447	148	542172	JX648369	
	Pina-M1-III	447	148	551064	JX648370	
Ae. markgrafii	*Pina-C1-I*	447	148	551132	JX648371	
				254863	JX648372	
				551119	JX648373	
	Pina-C1-II	447	148	203431	JX648374	
	Pina-C1-III	447	148	554237	JX648375	
Ae. searsii	*Pina-S^{s}1-I*	447	148	599151	JX648377	DQ269840 [2]
	Pina-S^{s}1-II	447	147	599174	JX648376	AY622792 [3]
Ae. speltoides	*Pina-S1-I*	447	148	393493	JX648378	DQ269833 [2]
	Pina-S1-II	444	147	573448	JX648379	
	Pina-S1-III	447	148	554296	JX648380	DQ269831, DQ269830, DQ269829 [2]
				486262	JX648381	
	Pina-S1-IV	447	148	554304	JX648382	
Ae. umbellulata	*Pina-U1-I*	447	-	554390	JX648383	AY608598 [4]
				542365	JX648384	
				222762	JX648385	
	Pina-U1-II	447	148	554409	JX648386	
				554417	JX648387	
	Pina-U1-III	447	148	542377	JX648388	
	Pina-U1-IV	447	148	560556	JX648389	DQ269847 [2]

[1] Cenci et al. (unpublished data); [2] Massa and Morris (2006); [3] Morris et al. (2001); and [4] Huo et al. (unpublished data).

Table 3. *Pinb* alleles in diploids species of the *Aegilops* genus.

Species	Allele	Size		Accession (PI)	NCBI accession	Previous NCBI accessions [references]
		DNA (bp)	Protein (aa)			
Ae. comosa	*Pinb-M1-I*	450	149	551031	JX648338	
				551017	JX648339	
	Pinb-M1-II	450	149	551053	JX648340	
	Pinb-M1-III	426	141	551038	JX648341	FJ898246 [1]
	Pinb-M1-IV	450	149	551022	JX648342	
	Pinb-M1-V	450	149	551064	JX648343	
Ae. markgrafii	*Pinb-C1-I*	447	148	573418	JX648344	
				564194	JX648345	
	Pinb-C1-II	447	148	542202	JX648346	
	Pinb-C1-III	447	148	551129	JX648347	
	Pinb-C1-IV	447	148	573413	JX648348	AY649747 (*Pinb-D1-i*) [2]; AY251995 [3]
Ae. searsii	*Pinb-S^{s}1-I*	444	147	599123	JX648349	AY622805 (*Pinb-S^{s}1-a*) [2]; DQ269874, DQ269873 [4]
	Pinb-S^{s}1-II	444	147	599175	JX648350	AY622806 (*Pinb-S^{s}1-b*) [2]; DQ269871, DQ269872 [4]
	Pinb-S^{s}1-III	447	148	599134	JX648351	AY649747 (*Pinb-D1-i*) [2]; AY251995 [3]
Ae. speltoides	*Pinb-S1-I*	444	147	487233	JX648352	FJ898251, FJ898257 [1]
	Pinb-S1-II	447	148	487236	JX648354	FJ898247, FJ898258, FJ898259 [1]
	Pinb-S1-III	447	148	486262	JX648355	
	Pinb-S1-IV	444	147	486262	JX648357	
	Pinb-S1-V	444	147	487232	JX648358	
	Pinb-S1-VI	444	147	487231	JX648359	FJ898250, FJ898256 [1]
	Pinb-S1-VII	444	147	554296	JX648353	AY622801 (*Pinb-S1-c*) [2]; DQ269863 [4]
	Pinb-S1-VIII	447	148	170203	JX648356	AY622803 (*Pinb-S1-e*) [2]; DQ269866 [4]
	Pinb-S1-IX	447	148	554304	JX648360	
Ae. umbellulata	*Pinb-U1-I*	450	149	564235	JX648361	
				554401	JX648362	
				542372	JX648363	
				487219	JX648364	
	Pinb-U1-II	447	148	542377	JX648365	
	Pinb-U1-III	447	148	Ciae66	JX648366	

[1] Cenci et al. (unpublished data); [2] Morris et al. (2001); [3] Massa et al. (2004); and [4] Massa and Morris (2006).

Regarding *Pinb*, the coding region ranged from 426 bp to 450 bp. This was due to 3-21 bp deletions and 3 bp insertions. For the *Pinb* gene, all sequenced accessions of *Ae. comosa* were 450 bp in size because they shared a 3 bp insertion at the N-terminal cleavable peptide, with the exception of *Pinb-M1-III*, which had a 21 bp deletion that

included part of the N-terminal cleavable peptide. This resulted in a sequence of 426 bp. One allele of *Ae. umbellulata* (*Pinb-U1-I*) also had the same 3 bp insertion seen in the *Ae. comosa* sequences. Two *Ae. searsii* sequences: *Pinb-S^{s}1-I* and *Pinb-S^{s}1-II*, showed a deletion of a triplet at the start of the N-terminal of the mature protein, which gave rise to a 444 bp sequence. Five *Ae. speltoides* alleles shared the same deletion (*Pinb-S1-I, Pinb-S1-IV, Pinb-S1-V, Pinb-S1-VI* and *Pinb-S1-VII*).

In *Pinb*, some alleles also showed similar DNA sequences with alleles previously reported in different species. The *Pinb-M1-III* allele of *Ae. comosa* was identical to one in *Ae. uniaristata* (NCBI: FJ898246; Cenci et al. unpublished data), the *Pinb-C1-IV* allele of *Ae. markgrafii* was the same as the allele described in *Ae. tauschii* (*Pinb-D1-i*; NCBI: AY649747) by Morris et al. (2001), and the *Pinb-S^{s}1-III* allele found in *Ae. searsii*. Finally, the *Pinb-S1-IX* allele found in *Ae. speltoides* was identical to the *Pinb-C1-III* of *Ae. markgrafii.*

Table 4. Identities among the nucleotide sequences (%).

		Interspecific			
Gene	Intraspecific	*Ae. comosa*	*Ae. markgrafii*	*Ae. searsii*	*Ae. speltoides*
Pina					
Ae. comosa	99.0				
Ae. markgrafii	98.6	98.6			
Ae. searsii	99.8	97.7	97.4		
Ae. speltoides	97.7	97.5	97.5	97.9	
Ae. umbellulata	98.4	98.3	98.2	97.5	97.6
Pinb					
Ae. comosa	96.5				
Ae. markgrafii	96.5	95.3			
Ae. searsii	93.6	93.0	94.2		
Ae. speltoides	96.8	93.7	94.8	96.2	
Ae. umbellulata	95.4	95.9	96.0	93.0	94.7

For *Pina*, intraspecific sequence identity in the coding regions ranged from 97.7 to 99.8% (Table 4). The species with the S genome had an identity of 98.7% and the species group with the C, M and U genomes had an identity of 98.7%. The overall coding sequence identity was 98.7%. In contrast, intra-specific and overall sequence identities in *Pinb* were 93.6–96.8% and 95.8%, respectively. The species group with the S genome had an identity of 95.2% and the species group with C, M and U genomes had an identity of 96.1%

Nucleotide diversity

For overall *Pina* sequences, 40 mutations at 36 polymorphic sites were detected (Table 5), of which 11 where synonymous and 29 were non-synonymous. Based on these single nucleotide polymorphisms (SNPs), 14 different haplotypes were found. The average number of nucleotide differences was 11. A larger number of polymorphic sites were found for the sequences in the group of species with the S genome than for the sequences in the group of species with the C, M and U genomes (24 and 22, respectively). This was also true for the average number of nucleotide differences (nine and eight, respectively).

For the *Pinb* sequences, the alignment of the overall coding sequences displayed a high level of SNPs with 83 mutations (39 synonymous and 44 non-synonymous) at 79 polymorphic sites (Table 5). On this basis, 22 different haplotypes were classified and an average number of 24 nucleotide differences between sequences were detected. As with *Pina*, a larger number of polymorphic sites were found in the group of species with the S genome than were found in the group of species with the C, M and U genomes (56 and 49, respectively). This was also true for the average number of nucleotide differences (16 and 14, respectively). However, these values were greater in *Pinb* than in *Pina*.

Table 5. Summary of DNA polymorphism and test statistics for selection in diploids species of the *Aegilops* genus.

Parameter	*Pina*			*Pinb*		
	Total	Group I*	Group II*	Total	Group I*	Group II*
N	16	10	6	24	12	12
η	40	24	25	83	50	59
k	10.93	7.89	9.47	24.29	14.41	15.86
s	36	22	24	79	49	56
h	14	9	6	22	12	12
$\theta \times 10^{-3}$	24.27	17.4	23.51	50.01	38.09	41.77
$\pi \times 10^{-3}$	24.44	17.65	21.18	57.41	33.82	35.73
D	-0.393 ns	-0.333 ns	-0.853 ns	0.366 ns	-0.595 ns	-0.867 ns

* Group I: species with genomes C, M and U; Group II: species with genome S.
η: total number of mutations; k: average number of nucleotide differences; s: number of polymorphic sites; h: number of haplotypes; θ: Watterson's estimate; π: nucleotide diversity; and D: Tajima's estimate D-test; ns: not significant.

A neutrality test was used to evaluate the degree of deviation from an equilibrium neutral model. For this test, two parameters were calculated: the number of segregating sites (θ) and the average number of nucleotide differences per site in two sequences (π). Under a drift-mutation balance, these two estimates are expected to give similar values;

otherwise, any form of natural selection may explain the maintenance of genetic variation (Table 5). The levels and patterns of polymorphism for the *Pina* and *Pinb* genes were consistent with a neutral equilibrium, according to Tajima's D test. Similarly, the results of both groups of sequences (the group of species with the S genome and the group of species with the C, M and U genomes) corresponding to each gene were found to be in neutral equilibrium.

Pina and Pinb deduced proteins

The deduced protein sequences from the *Pin* alleles found in this study were compared with *Pina-D1a* and *Pina-D1b* deduced protein sequences (Fig. 3 and 4). The protein size was more homogenous in PINA than in PINB. All the PINA sequences had 148 residues, with the exception of PINA derived from the *Pina-S^{s}1-II* allele of *Ae. searsii*, which had 147 residues due to a substitution at residue 148 where one tryptophan had been changed to a stop codon (Table 2). For PINB, the sequence lengths were 149, 148, 147 and 141 amino acids (Table 3).

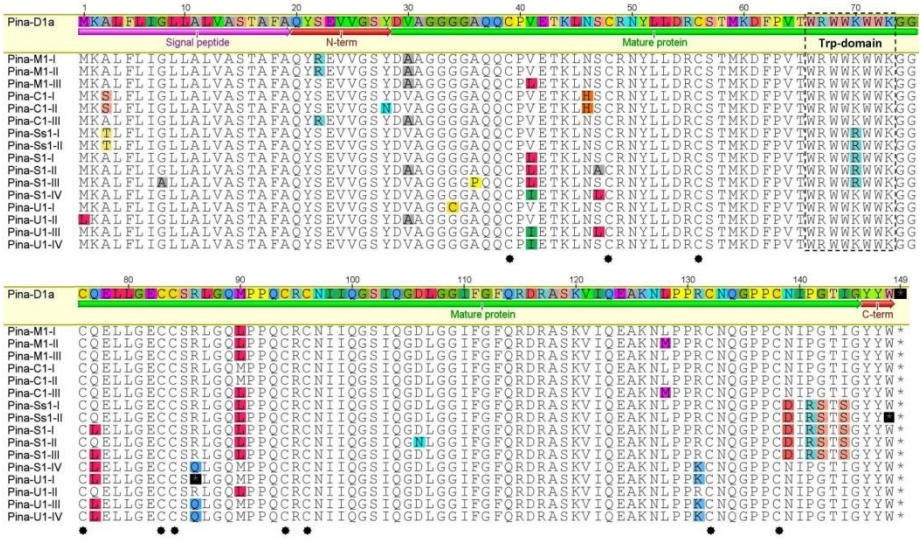


Figure 3. Alignment of deduced amino acid sequences of PINA from alleles described in this study and bread wheat cv. Chinese Spring (*Pina-D1a*). The ten cysteine residues are marked with asterisks. Tryptophan-rich residue is indicated by box.

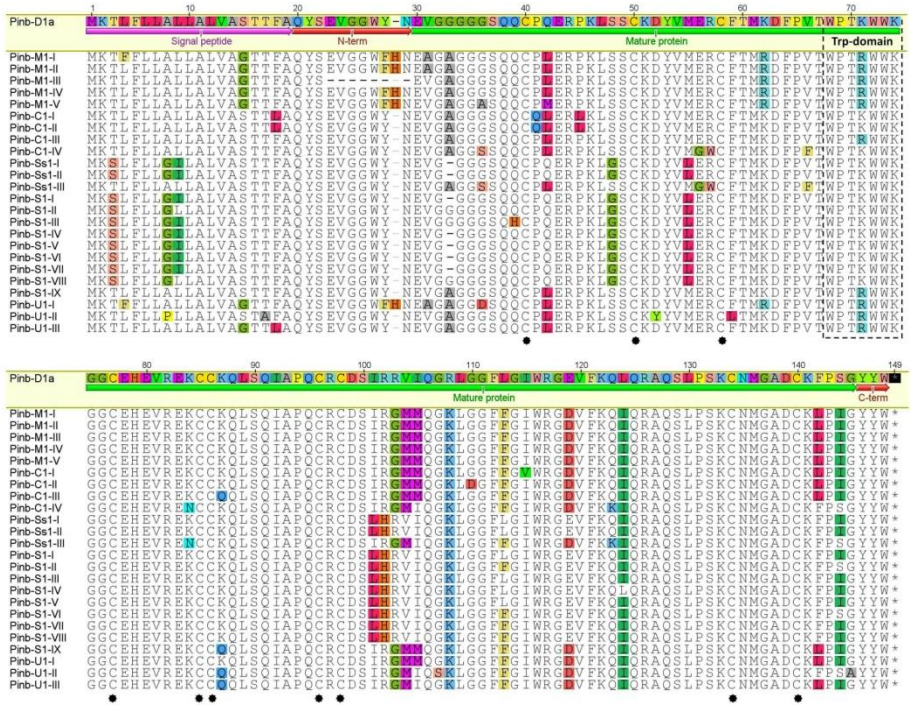


Figure 4. Alignment of deduced amino acid sequences of PINB from alleles described in this study and bread wheat cv. Chinese Spring (*Pinb-D1a*). The ten cysteine residues are marked with asterisks. Tryptophan-rich residue is indicated by box.

For PINA, the number of polypeptides within each species was similar to the number of *Pina* alleles found (16). However, four polypeptides were the same, although they appeared in different species. The deduced polypeptide of the *Pina-C1-III* allele (*Ae. markgrafii*) was identical to the *Pina-M1-II* allele (*Ae. comosa*) and both species were from Group I. The same situation was detected between the *Pina-S1-IV* allele (*Ae. speltoides*) and the *Pina-U1-III* allele (*Ae. umbellulata*), although in this case, each species was from a different group. The 24 *Pinb* alleles were translated into 22 polypeptides, since the *Pinb-S^{s}1-I* and *Pinb-S^{s}1-II* alleles (*Ae. searsii*) on the one hand and the *Pinb-S1-I* and *Pinb-S1V* alleles (*Ae. speltoides*) on the other had the same amino acid sequences.

Both PIN proteins are synthesised as precursors or pre-proteins and contained four domains: a signal peptide, a N-terminal, a mature protein and a C-terminal (Gautier et al. 1994). The differences within each domain were analysed in each *Aegilops* sequence with respect to the reference sequences (PINA and PINB from cv. Chinese Spring). The less variable domain was clearly the C-terminal, where only one change was found, and the

Pina-S^{s}1-II allele showed one change (Trp148 → stop codon) in the third amino acid residue of this domain, which implied that the PINA precursor that coded this allele had 147 residues instead of 148 (Fig. 3).

Of the other two cleavable domains (the signal peptide and N-terminal), the most variable was the signal peptide. Four and eight changes were detected in PINA and PINB, respectively (Fig. 3 and 4). One of the most remarkable changes was the change: Met → Leu in the first residue, and this was only detected in the *Pina-U1-II* allele of *Ae. umbellulata* (Fig. 3), which was confirmed in both *Ae. umbellulata* accessions that showed this allele.

The changes in the N-terminal domain were minor. In PINA, only two changes were detected: Ser22 → Arg in three alleles (*Pina-M1-I*, *Pina-M1-II* and *Pina-C1-III*) and Tyr28 → Asn in the *Pina-C1-II* allele. Four alleles in *Ae. comosa* (*Pinb-M1-I*, *Pinb-M1-II*, *Pinb-M1-IV* and *Pinb-M1-V*) and one in *Ae. umbellulata* (*Pinb-U1-I*) showed the change: Tyr28 → Phe, together with an extra histidine between positions 28 and 29 of the reference sequence (PINB from CS). The *Pinb-M1-III* allele lacked eight amino acids compared with the total *Ae. comosa* sequence, which represented almost all of the N-terminal cleavable peptide (Fig. 3 and 4).

Concerning the mature proteins, the two main features of the puroindolines were almost completely conserved among all the analyzed sequences. Both PINA and PINB had a cysteine backbone made up of 10 cysteines. It contained a Cys-Cys pair, a triplet Cys-X-Cys and a tryptophan-rich domain. According to Feiz et al. (2009), the mature protein is divided into three functional regions: region-1, the beginning of the mature protein up to the third cysteine residue; region-2, between the third and fifth cysteine residue, which includes the tryptophan-rich domain, and region-3, the fifth cysteine residue up the end of the mature protein.

This study showed that up to 20 and 37 substitutions were detected in the mature protein of PINA and PINB sequences, respectively. In both proteins, the most variable regions were: region-1, with eight changes for PINA and 15 changes for PINB, and region-3, with 11 changes for PINA and 17 for PINB (Fig. 3 and 4). Most of these changes were conservative, which suggested that their influence on protein properties would be low. In PINA, three positions in region-1 showed non-conservative substitutions (Gly34 → Cys, Ala36 → Pro and Ser47 → Ala/Leu), whereas in PINB, two *Ae. searsii* sequences (*Pinb-S^{s}1-I* and *Pinb-S^{s}1-II*) lacked an amino acid at position 3 of

the N-terminal extreme in the mature protein and this gave rise to a protein containing 147 amino acids. Five *Ae. speltoides* alleles shared the same deletion (*Pinb-S1-I, Pinb-S1-IV, Pinb-S1-V, Pinb-S1-VI* and *Pinb-S1-VII*). Another remarkable change was the change: Arg57 → Trp, detected in the *Pinb-C1-IV* allele (*Ae. markgrafii*) and the *Pinb-S^{s}1-III* allele (*Ae. searsii*). Inside region-3, the most important change was detected in the *Pina-U-I* allele and was a change from Arg86 to a premature termination codon in the mature protein, which probably results in a truncated protein (Fig. 3 and 4).

Within the materials evaluated here, some sequences in PINA and PINB contained the Lys70 → Arg or Lys71 → Arg substitution, respectively, which occurred in the tryptophan-rich domain found in region-2. For PINA the substitution was seen in all sequences of *Ae. searsii* and *Ae. speltoides*, with the exception of the *Pina-S1-IV* allele, whereas in PINB this change occurred in all versions of *Ae. comosa* and *Ae. umbellulata* and in one allele from *Ae. markgrafii* (*Pinb-C1-III*) and one from *Ae. speltoides* (*Pinb-S1-IX*). One other replacement was observed in region-2 of PINA (Gln77 → Leu) and five were observed in PINB (Phe59 → Leu, Lys62 → Arg, Val66 → Phe and Lys84 → Asn).

Discussion

In this study, *Pina* and *Pinb* gene variability at the molecular level has been examined in five species of the *Aegilops* genus. The results showed that there were two different patterns of variation, and a greater level of nucleotide polymorphism was found for *Pinb* than for *Pina*, which corresponded with the larger number of *Pinb* alleles reported for bread wheat (for a review, see Morris y Bhave 2008). This polymorphism resulted in 16 alleles for the *Pina* gene and 24 alleles for the *Pinb* gene. Ten and 17 of them, respectively, were novel. These results were consistent with a neutral equilibrium by the Tajima test, although the greater polymorphism levels and substitution values detected for *Pinb* compared to *Pina* confirmed that the *Pina* gene was less variable within the species evaluated. This could be related to the suggestion that the *Pina* gene is more important than *Pinb* in plant defence (Krishnamurthy et al. 2001), which has led to the fixation of adaptive mutations and a reduction in variation.

A large number of changes, compared to the wild-type alleles of bread wheat, were observed in the alleles found in this study. The first novel outstanding mutation detected was the change in the start codon (ATG for TTG) in the *Pina-U1-II* allele. This kind of change (non-AUG initiated translation) is extremely rare in eukaryotic cells and

few genes carrying this mutation have been described in plants (Riechman et al. 1999; Christensen et al. 2005; Kobayashi et al. 2002). Gordon et al. (1992) showed that non-AUG codons could lead to an efficient initiation of the translation of a chloramphenicol acetyl transferase (CAT) in plant cells, although the expression of the gene was reduced. The level of the expression reduction (from 70 to 100%) depended on the codon that substituted AUG and the adjacent sequences. In this study, AUG was replaced by UUG, and based on previous results, this could lead to a PINA reduction of 97%, depending on the adjacent sequences. This fact could cause a new intermediate hardness level between soft and hard.

Several studies have shown that the conservation of specific regions and domains present in PINA and PINB proteins are essential if the softness of the grain texture is to be preserved (Corona et al. 2001; Feiz et al. 2009). In this respect, the cysteine backbone is critical for the stabilization of the structure, whereas the tryptophan-rich domain acts as the binding site with the amyloplast lipid bilayer membrane (Douliez et al. 2000). In the sequences analysed in the current study, no change was detected in the cysteine backbone, with the exception of *Pina-U1-I*, which had one additional cysteine residue at position 34. However, it was not possible to evaluate the effect of this change on the PINA protein, because this allele also showed a premature stop codon that probably leads to the production of a truncated protein.

Another important issue is that the PINs are synthesized as precursors. These precursors consist of a signal peptide, two cleavable domains (N-terminal and C-terminal) and the mature protein (Gautier et al. 1994). These three domains have important functions during the processing of the mature protein. Consequently, any mutation in these regions could have an effect on grain endosperm texture due to the correct or incorrect processing of the PIN precursors. In the current study, changes were found in these three regions in both PINA and PINB. For example, a change in the *Pina-S^s1-II* allele removed the last tryptophan residue in the PINA C-terminus. The same mutation for PINA in S genome species was described by Lillemo et al. (2002) and Simeone et al. (2006). The hydrophobic residues, such as the tryptophan found in the C-terminal, are known to be important in the translocation of PINA through the membranes. The reduced levels of PINA with the same change in starch granules (Gazza et al. 2006) suggested that a mutation in this zone could alter the translocation of PINA to the starchy endosperm. In the current study, a novel Tys-28 $\rightarrow$ Asn mutation was observed at the N-terminal cleavage peptide for PINA in *Pina-C1-II*. This substitution could affect the

cleavage of this peptide because the elimination of this region occurred between positions 28 and 29. Furthermore, a deletion in the N-terminal domain was detected for PINB in *Pinb-M-III*. The deletion, spanning 21 bp, which led to the loss of seven amino acids, was close to the cleavage site of the N-terminal peptide. This deletion has only been detected in the $A^mA^mS^{sh}S^{sh}$ amphiploid by Li et al. (2008a) and one accession of *Ae. uniaristata*. However, no studies about its effect on grain hardness have been undertaken. A mutation detected in the N-terminal domain of PINB occurred in most *Ae. comosa* accessions and in one *Ae. umbellulata* accession (*Pinb-U1-I* allele). The mutation was an insertion of one His residue between positions 28 and 29 and could lead to the incorrect cleavage of this domain from the precursor during the formation of the mature protein, which would alter the function of the protein.

Other novel changes, such as Gly34 → Cys and Leu-128 → Met for PINA and Gly36 → Asp, Gln-39 → His and Ile-115 → Val for PINB, were detected in the mature protein in addition to the mutations defined within the specific cleavage regions and domains of PINA and PINB. These changes, with respect to wild-type alleles, could also alter grain hardness.

Many studies have shown that mutations within the tryptophan-rich domain or in the adjacent residues result in a hard texture (Giroux and Morris 1997, 1998; Lillemo and Morris 2000), together with very low levels of PIN protein on the surface of starch granules (Corona et al. 2001). Feiz et al. (2009) established three functional regions within the mature protein, which included the region between the third and fifth cysteine residues containing the tryptophan-rich domain where the amino acid changes have significant effects on grain hardness. A conservative change was observed in PINA in the tryptophan--rich domain (Lys70 → Arg) for almost all members of the S-genome group (*Ae. searsii* and *Ae. speltoides*), with the exception of the *Pina-S1-IV* allele, and this confirmed the results of previous researches (Massa and Morris 2006; Simeone et al. 2006). The same mutation in PINB (Lys71 → Arg) was mainly found in *Ae. comosa* and *Ae. umbellulata*. Simeone et al. (2006) suggested that this mutation in *Pina* of the S-genome group has no influence on grain hardness. Similar results were found in the *Pina-N1a* allele of *Ae. ventricosa* Tausch. by Gazza et al. (2006), who indicated that this change does not affect the amount of puroindoline accumulated on the starch granules. Moreover, several substitutions were observed near the tryptophan-rich domain of both PINA and PINB, such as Arg86 → Gln in PINA and Arg-57 → Trp in PINB. These

mutations, when present in synthetic wheats, led to a softer phenotype than the wild type alleles (Gedye et al. 2004) and could be used to modify soft wheat milling properties (Reynolds et al. 2010a,b).

In conclusion, the large number of novel *Pina* and *Pinb* alleles identified in this study shows that the diploid species of the genus *Aegilops* could be a good source of genetic variability for both *Pina* and *Pinb* genes, and this could be used in breeding programmes to extend the range of textures in wheat. Additional studies are required in order to investigate these alleles further and to evaluate their effect on modern wheat grain hardness.

Acknowledgements

This research was supported by grant AGL2010-19643-C02-01 from the Spanish Ministry of Economy and Competitiveness, and the European Regional Development Fund (FEDER) from the European Union. The first author is grateful to the Spanish Ministry of Economy and Competitiveness (FPI programme) and European Social Fund for a predoctoral fellowship.

Supplementary material

Supplementary Table 1. Plant material used to survey *Puroindoline* gene sequence in diploids species of the *Aegilops* genus.

Accession	Country	*Pina*	*Pinb*
Ae. comosa (genome M)			
PI 542172	Turkey	II	I
PI 542174	Turkey	I	IV
PI 542176	Turkey	I	IV
PI 551017	Greece	I	I
PI 551018	Greece	I	I
PI 551024	Greece	I	IV
PI 551029	Greece	I	NULL
PI 551031	Greece	I	I
PI 551036	Greece	I	NULL
PI 551022	Greece	I	IV
PI 551038	Greece	II	III
PI 551039	Greece	I	II
PI 551042	Greece	I	II/III
PI 551043	Greece	I	II
PI 551048	Greece	I	II
PI 551053	Greece	I	II
PI 551064	Greece	III	V
PI 551076	Greece	II	NULL
Ae. markgrafii (genome C)			
PI 203431	Turkey	II	II
PI 254863	Iraq	I	II
PI 263554	Turkey	II	II
PI 542199	Turkey	II	II
PI 542202	Turkey	II	II
PI 542208	Turkey	II	I
PÌ 542209	Turkey	I	I
PI 551119	Greece	I	IV
PI 551122	Greece	II	I
PI 551129	Greece	I	III
PI 551132	Greece	I	IV
PI 551147	Greece	I	IV
PI 554237	Turkey	II	III
PI 564194	Turkey	II	I
PI 573413	Turkey	I	IV
PI 573415	Turkey	II	I
PI 573418	Turkey	I	I
Ae. searsii (genome S^s)			
PI 599122	Israel	II	I
PI 599123	Israel	II	I
PI 599130	Jordan	II	II
PI 599134	Jordan	I	II/III
PI 599149	Israel	I	II/III
PI 599150	Israel	II	II
PI 599151	Israel	I	II

Accession	Country	*Pina*	*Pinb*
PI 599152	Israel	I	I
PI 599156	Israel	I	I
PI 599158	Israel	II	II
PI 599163	Israel	II	II
PI 599164	Israel	I	II
PI 599168	Israel	II	II
PI 599174	Israel	II	I
PI 599175	Israel	I	II
Ae. speltoides (genome S)			
PI 170203	Turkey	III	VIII
PI 219867	Iraq	II	VIII
PI 393493	Israel	I	VIII
PI 486262	Turkey	III	IV/III
PI 486263	Turkey	II	I
PI 487231	Syria	I	VI
PI 487232	Syria	I	V
PI 487233	Syria	I	I
PI 487236	Syria	II	II
PI 487238	Syria	II	VI
PI 499261	China	I	VIII
PI 554295	Turkey	I	VIII
PI 554296	Turkey	III	VII
PI 554298	Turkey	II	III
PI 554304	Turkey	IV	IX
PI 573448	Turkey	II	VII
Ae. umbellulata (genome U)			
Ciae 66	Yugoslavia	I	III
PI 222762	Iran	I	NULL
PI 298906	Iraq	I	I
PI 428569	Azerbaijan	III	I
PI 487219	Syria	III	I
PI 542362	Turkey	III	I
PI 542365	Turkey	I	NULL
PI 542372	Turkey	III	I
PI 542377	Turkey	III	II
PI 542381	Turkey	III	II
PI 542383	Turkey	III	II
PI 554388	Turkey	I	I
PI 554390	Turkey	I	NULL
PI 554401	Turkey	III	I
PI 554409	Turkey	II	III
PI 554417	Turkey	II	III
PI 560556	Turkey	IV	I
PI 564235	Turkey	III	I

CAPÍTULO II

CARACTERIZACIÓN Y DIVERSIDAD GENÉTICA DEL GEN *Gsp-1* EN ESPECIES DIPLOIDES DEL GÉNERO *Aegilops*

Publicado como:

S. Cuesta, J.B. Alvarez, C. Guzmán (2015) Characterization and sequence diversity of the *Gsp-1* gene in diploid species of the *Aegilops* genus. *Journal of Cereal Science* **63:** 1-7.

Resumen

Algunos estudios han sugerido un papel menor en la determinación de la textura del endospermo para el gen *Gsp-1*, aunque parece estar relacionado con la síntesis del péptido arabinogalactano. El presente estudio evaluó la variabilidad alélica de los genes *Gsp-1* en cinco especies diploides del género *Aegilops*, junto con la caracterización molecular de las principales variantes alélicas encontradas en cada especie y sus relaciones filogenéticas con los trigos poliploides. Se encontró un alto nivel de polimorfimso en los genes *Gsp-1* – hasta 19 alelos fueron detectados, de los cuales 11 fueron nuevos. El procesamiento de las secuencias aminoacídicas de *Gsp-1* dio lugar a dos principales dominios: AGP (péptido arabinogalactano) y GSP-1 (proteína de la suavidad del grano). Algunas sustituciones fueron detectadas en ambos dominios con respecto a las secuencias del trigo harinero, las cuales podrían modificar la dureza del grano. Las secuencias obtenidas mostraron relación con los genes *Gsp-B1* y *Gsp-D1* del trigo común. Estos resultados mostraron que las especies diploides de *Aegilops* son una rica fuente de nueva variabilidad genética para el gen *Gsp-1*, cuya funcionalidad hipotética debería ser analizada dentro del acervo genético del trigo moderno, valorando el verdadero papel de estos genes en la textura del grano y su papel como fuente del AGP, un polisacárido no almidonado con propiedades saludables.

Palabras clave: péptido arabinogalactano, dureza del grano, proteína de la suavidad del grano, trigo.

Abstract

Several studies have suggested a minor role in determining grain endosperm texture for the *Gsp-1* gene, although appears related with the synthesis of the arabinogalactan-peptide. The current study evaluated the allelic variability of *Gsp-1* genes in five diploid species of the *Aegilops* genus along with the molecular characterization of the main allelic variants found in each species and their phylogenetic relationships with the polyploid wheats. There was a high level of polymorphism in *Gsp-1* genes - up to 19 alleles were obtained, of which 11 were novel. The processing of *Gsp-1* amino acid sequences led to two main domains: AGP (arabinogalactan peptide) and GSP-1 (grain softness protein). Several substitutions were detected in both domains with respect to the bread wheat sequences, which could modify grain hardness. The sequences obtained showed relationship with the *Gsp-B1* and *Gsp-D1* genes of common wheat. These results showed that diploid *Aegilops* species are a rich resource of novel genetic variability for *Gsp-1* gene, whose hypothetical functionality should be tested in the genetic pool of modern wheat, valuing as the true role of these genes in the grain hardness and their effective role as source of the AGP, a non-starch polysaccharide with healthy properties.

Key words: Arabinogalactan peptide, grain hardness, grain softness proteins, wheat.

Introduction

Grain hardness, the most important trait to define wheat end-use quality, has been associated with a ~13 kDa complex of proteins, formed by two major proteins, Puroindoline a (PINA) and Puroindoline b (PINB), together with a minor component: Grain Softness Protein-1 (GSP-1). These proteins are synthesized by genes (*Pina*, *Pinb* and *Gsp-1*, respectively) located at the *Ha* locus. This locus occurs on the group 5 chromosomes of all *Poaceae* species (Wilkinson et al. 2013), but in durum wheat (*Triticum turgidum* ssp. *durum* Desf. em. Husn.; 2n = 4× = 28, AABB) and bread wheat (*T. aestivum* L. ssp. *aestivum*; 2n = 6× = 42, AABBDD) a partial deletion has been detected on the orthologous loci of chromosomes 5A and 5B. This deletion implies that both *Pin* genes are missing but not the *Gsp-1* gene (Li et al. 2008a). This event has generated changes in the grain texture of polyploid wheat, because durum wheat that lacks both *Pin* genes shows a very hard texture, whereas bread wheat, which has the *Pin* genes of chromosome 5D derived from *Aegilops tauschii* Coss. (2n = 2× = 14, DD), has a hard or soft texture.

Although the main determinant of grain texture has been associated with the presence/absence or modification of *Pin* genes (Bhave and Morris 2008b; Feiz et al. 2009), some studies have suggested that the *Gsp-1* gene could also have a role in this important grain trait (Morris et al. 2013). Although Tranquilli et al. (2002) found that deletions or allelic variants of *Gsp-A1* and *Gsp-B1* had no significant impact on grain texture, Gedye et al. (2004) showed that a significant amount of variation in hardness of synthetic wheat was assignable to durum wheat parent and suggested a role for GSP-1 and other QTLs associated with the hardness. Additional data that could support this hypothesis is the grain texture variation detected in durum wheat, which has no *Pin* genes, or in bread wheat lines showing the same *Pin* alleles composition (Bhave and Morris 2008b).

The *Gsp-1* gene has an intronless sequence of ~495 nucleotides with high similarity to the *Pin* genes, which supports the hypothesis of the role of this gene in grain hardness. Similar to those of PINA and PINB, the GSP-1 protein exhibits a tryptophan (Trp)-rich domain and a cysteine backbone, formed by 10 cysteine residues. The main difference between *Gsp-1* and *Pin* genes is the presence of one domain of 15 residues in

the N-terminal region of *Gsp-1* after the peptide signal. This domain has been associated with the arabinogalactan peptide (AGP), which is part of the non-starch polysaccharides synthesized in the cell walls and located in plasma membranes, such as the amyloplast membrane, and that could affect the strength adhesion of starch granules and surrounding protein matrix (Van den Bulck et al. 2002). In fact, Bettge and Morris (2000) showed that in soft wheat up to 76% of the variation in hardness was caused by non-starchy polysaccharides. Some studies have suggested that *Gsp-1* leads to the synthesis of a 164-amino-acid preproprotein, which undergoes a posttranslational maturation with at least three proteolytic cleavages (Wilkinson et al. 2013). Using the *Gsp-D1* gene as standard, the first proteolytic cleavage eliminates the peptide signal (Met1–Ala19), while the second cleavage gives a peptide of 29 residues (Gln20-Ser48) that is processed into two different peptides: 15-mer putative AGP (Tyr21-Asp35) and 13-mer peptide (Gly36-Ser48). This peptide is shorter in PINA and PINB, forming part of their N-terminal domains. One last cleavage at Gly161 leads to the release of the sequence (Gly49-Gly161) of the mature GSP-1 protein (Elmorjani et al. 2013; Wilkinson et al. 2013).

One important source of gene variation for modern wheat breeding is the relative species, whose genomes are related with the genomes present in wheat. These species showed large variability in interesting traits such as disease resistance, stress tolerance and quality (Scheiner et al. 2008). In particular, *Aegilops* species are good candidates for novel variability that can be used to enlarge the genetic pool available for breeders. Species with the S genome (*Ae. searsii* Feldman & Kislev ex K. Hammer or *Ae. speltoides* Tausch) would be related to the B genome; whereas species such as *Ae. comosa* Sm., *Ae. markgrafii* (Greuter) K. Hammer or *Ae. umbellulata* Zhuk., which possess the M, C or U genomes are related along with *Ae. tauschii* to the D genome. In a previous study (Cuesta et al. 2013), we used one collection of these species to analyse the genetic variability for *Pina-1* and *Pinb-1*, finding 16 and 24 alleles, respectively. The combination of both genes revealed up to 42 different genotypes, which also seem good candidates to search for novel variability for *Gsp-1*.

The current study attempted to evaluate the allelic variability of *Gsp-1* in the 42 accessions from five diploid *Aegilops* species previously analysed for *Pin* genes, along with the molecular characterization of the main allelic variants found in each species and their phylogenetic relationships with the polyploid wheats.

Materials and Methods

Plant vegetal

Forty-two accessions of five diploids species of the *Aegilops* genus: *Ae. comosa*, *Ae. markgrafii*, *Ae. searsii*, *Ae. speltoides* and *Ae. umbellulata*, which were previously analysed for *Pin* genes by Cuesta et al. (2013), were used in this study (Supplementary Table 1). These materials were obtained from the National Small Grains Collection (Aberdeen, ID, USA).

PCR amplification and sequencing of Gsp-1 gene

Genomic DNA was isolated from young leaves of a single plant per accession according to the method of CTAB (Stacey and Isaac 1994). The *Gsp-1* gene was amplified with the primers 5′-GGTGTGGCCTCATCTCATCT-3′ and 5′-AAATGGAAGCTACATCACCAGT-3′ designed by Massa et al. (2004). Reactions were performed in 15 μl of total volume containing 50 ng of genomic DNA, 0.4 μM of each primer, 0.2 mM of dNTPs, 1.5 mM of $MgCl_2$, 1× of reaction buffer and 0.75 U of *Taq* DNA polymerase (Promega, Madison, WI, USA). The amplification was performed with an initial denaturation step of 3 min at 94 °C followed by 35 cycles as follows: 45 s at 94 °C, a step of annealing of 30 s at 56 °C or 58 °C, then 30 s at 72 °C. To finish, a final extension step at 72 °C for 5 min was performed. The PCR products were separated by electrophoresis on polyacrylamide gels of 8% (p/v; C: 1.28%), stained with ethidium bromide and visualized under UV light.

PCR products were purified and ligated into pGEM-T easy vector (Promega, Madison, WI, USA) and cloned in *Escherichia coli* JM109 competent cells. Three positive clones for each PCR product were sequenced with M13 universal primers using an ABI Prism 310 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The novel sequences are available from GenBank database.

Data analysis

The sequences obtained were analyzed and compared using the Geneious Pro ver. 5.0.3 software (Biomatters Ltd.). DNA sequences of allelic wild-type *Gsp-B1* and *Gsp-D1* of common wheat cv. Renan (CR626930 and CR626934, respectively), were included for comparison (Chantret et al. 2005). DNA analyses were conducted by DNAsp ver. 5.0 (Librado and Rozas 2009) and parameters as total number of mutations (η), average

number of nucleotide differences (*k*), and number of polymorphic sites (*s*), were calculated. Nucleotide diversity was estimated as theta (θ), the number of segregating (polymorphic) sites (Watterson 1975), and pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987). Tests of neutrality were performed using Tajima's *D* statistic (1989).

Protein sequences were compared by Geneious Pro ver. 5.0.4 software (Biomatters Ltd.) and the Protein Logo software was used to create the logo representations of the deduced amino acid sequences of the *Gsp-1* gene. The overall height of the stack indicates the sequences conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position.

A phylogenetic tree was constructed with the software MEGA5 (Tamura et al. 2011) using the complete coding sequences obtained together with the sequences of the *Gsp-1* genes of common wheat cvs. Renan (*Gsp-A1*: CR626929; *Gsp-B1*: CR6226930; and *Gsp-D1*: CR626934) and Cranbrook (*Gsp-A1*: AY945214; *Gsp-B1*: AY945217; and *Gsp-D1*: AY945222), einkorn wheat - *T. monococcum* L. ssp. *monococcum* (*Gsp-A^{m}1*: AY491681) and *Ae. tauschii* (*Gsp-D1*: AF177219). A neighbour-joining cluster with all sequences analyzed was generated using the maximum composite likelihood method and one bootstrap consensus from 1000 replicates.

Results

Nucleotide variation of Gsp-1 genes

Based on the mobility of the PCR products (amplicons) obtained, the *Gsp-1* gene showed 2-7 variants in different species (Fig. 1). The *Pina* gene was amplified together with *Gsp-1* gene due to the high homology between them. The *Gsp-1* amplicons were sequenced on at least one representative accession, obtaining up to 19 alleles, with 11 of them novel (Table 1).

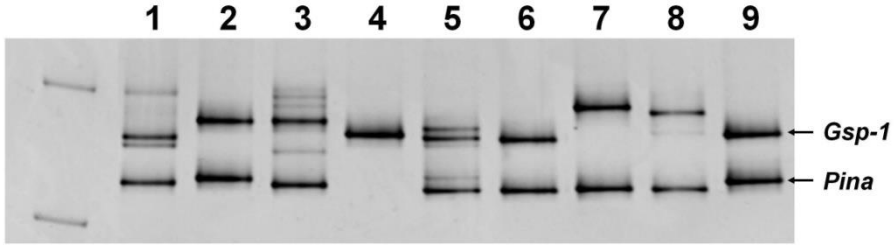


Figure 1. Representative sample of the allelic variation for the amplicons of *Gsp-1* gene detected in the *Aegilops* species evaluated. Lanes are as follow: 1, cv. Chinese Spring; 2 and 3, *Ae. comosa*; 4, *Ae. markgrafii*; 5 and 6, *Ae. speltoides*; 7 and 8, *Ae. searsii*; and 9, *Ae. umbellulata*.

Table 1. *Gsp-1* alleles in diploids species of the *Aegilops* genus.

Species	Allele	Size		Accession (PI)	NCBI accession	Previous NCBI accessions [references]
		DNA (bp)	Protein (aa)			
Ae. comosa	*Gsp-M1-I*	495	164	551064	KF861942	DQ269912 [2],
				551038	KF861961	DQ269914 [2],
						FJ898184[1]
	Gsp-M1-II	495	164	551042	KF861943	DQ269913 [2],
						FJ898189 [1],
						FJ898190 [1]
Ae. markgrafii	*Gsp-C1-I*	495	164	203431	KF861963	DQ269916 [2]
				564194	KF861962	
				573413	KF861944	
	Gsp-C1-II	495	164	203431	KF861945	
	Gsp-C1-III	495	164	554237	KF861946	DQ269913 [2],
						FJ898189[1], FJ
						898190[1]
	Gsp-C1-IV	495	164	573413	KF861947	
Ae. searsii	*Gsp-S^{s}1-I*	495	164	599134	KF861948	
	Gsp-S^{s}1-II	492	163	599152	KF861949	DQ269906[2],
						DQ269904[2]
Ae. speltoides	*Gsp-S1-I*	495	164	487233	KF861950	DQ269896 [2]
	Gsp-S1-II	495	164	393493	KF861951	EU307557 [1]
	Gsp-S1-III	492	163	486262	KF861952	DQ269900 [2],
						EU307555[1],
						EU307563[1]
	Gsp-S1-IV	492	163	554296	KF861953	
	Gsp-S1-V	495	164	554304	KF861954	
	Gsp-S1-VI	495	164	554304	KF861955	
	Gsp-S1-VII	492	163	573448	KF861956	
Ae. umbellulata	*Gsp-U1-I*	495	164	CIae 66	KF861957	
	Gsp-U1-II	495	164	542365	KF861958	
				542372	KF861964	
				542377	KF861965	
	Gsp-U1-III	495	164	554417	KF861959	
	Gsp-UI-IV	495	164	560556	KF861960	

[1] Cenci et al. (unpublished data); [2] Massa and Morris (2006).

The nucleotide sequences of the alleles detected were compared with *Gsp-1* sequences from B and D genomes (cv. Renan, common wheat). According to the above mentioned phylogenetic relationships, we grouped these species in two sets: Group I included species with C, M and U genomes, and Group II those with the S genome. Consequently, sequences from *Ae. searsii* and *Ae. speltoides* were aligned with *Gsp-B1* sequence, whereas alleles of *Ae. comosa*, *Ae. markgrafii* and *Ae. umbellulata* were aligned with *Gsp-D1* (Supplementary Figs. 1 and 2, respectively). The size in all cases was 495 bp, with the exception of four alleles (*Gsp-S^{s}1-II*, *Gsp-S1-III*, *Gsp-S1-IV* and *Gsp-S1-VII*) whose sequences were 492 bp, due to a deletion of one codon at position 106-108 (Table 1). Although, in general, the alleles from different species showed differences in their nucleotide sequences, some showed complete identity between them (e.g. *Gsp-C1-III* vs. *Gsp-M1-II*, or *Gsp-C1-II* vs. *Gsp-S1-V*), which could be understood as the consequence of convergent evolution (in the original alleles, the same mutations took place in different species to converge in the same allele). The presence of the described variability in the common ancestor of the different *Aegilops* species could be also a possible explanation to this finding.

The DNA polymorphism found in *Gsp-1* genes is indicated in Table 2. The sequences for Group I showed 43 polymorphism sites with 45 mutations, of which 18 were synonymous and 27 non-synonymous changes, with the average number of nucleotide differences of 17.40. The *Gsp-1* gene of Group II displayed the highest level of polymorphism with 53 polymorphism sites, although the average number of nucleotide differences was lower (16.70) than that of Group I. Among the 56 mutations detected, 30 were synonymous changes and 26 were non-synonymous. In both groups, the polymorphism detected resulted in nine different haplotypes.

For overall *Gsp-1* sequences (Groups I and II), the total number of mutations and average number of nucleotide differences were higher than those of individual groups, with 79 and 20.10, respectively; as were the number of polymorphic sites and haplotypes, with 71 and 17, respectively.

Two statistics, π and θ, were used to estimate nucleotide diversity (Table 2). Under a drift-mutation balance, these two estimates are expected to give similar values as was the case in the present study. Tajima's D-test was not significant either for both groups of sequences or overall sequences, which was consistent with a neutral equilibrium.

Table 2. Summary of DNA polymorphism and test statistics for selection in diploids species of the *Aegilops* genus.

Parameter	Group I*	Group II*	Total
N	10	9	19
η	45	56	79
k	17.40	16.70	20.11
s	43	53	71
h	9	9	17
$\theta \times 10^{-3}$	30.71	39.64	41.29
$\pi \times 10^{-3}$	35.15	33.93	40.86
D	0.457 ns	0.970 ns	-0.456 ns

* Group I: species with genomes C, M and U; Group II: species with genome S.
η: total number of mutations; *k*: average number of nucleotide differences; *s*: number of polymorphic sites; *h*: number of haplotypes; θ: Watterson's estimate; π: nucleotide diversity; and *D:* Tajima's estimate *D*-test; ns: not significant.

Amino acid sequences

The deduced amino acid sequence of these genes was formed by five domains (signal peptide, AGP, N-terminal, GSP-1 and C-terminal), which are processed during the maturation of this protein. Of these, only the AGP and GSP-1 domains have some associated function. The substitutions within the functional or non-functional regions were analysed with respect to the standard sequence in each case. The size was 164 amino acids for all alleles (Fig. 2a and Fig. 2b), with the exception of *Gsp-S^{s}1-II*, *Gsp-S1-III*, *Gsp-S1-IV* and *Gsp-S1-VII* alleles that showed one least amino acid inside the N-terminal domain.

Five positions inside the signal peptide showed variation, with up to seven different substitutions in overall sequences. In Group I, alleles *Gsp-U1-II* and *Gsp-U1-IV* showed one non-conservative substitution (Ala17 → Thr) (Fig. 2b). In Group II (Fig. 2a), there were three non-conservative substitutions: Thr3 → Ile for the *Gsp-S1-II* allele, Thr3 → Ala for *Gsp-S^{s}1-II* and *Gsp-S1-IV* alleles, and Thr17 → Ala for four alleles of *Ae. speltoides* (*Gsp-S1-III*, *Gsp-S1-IV*, *Gsp-S1-V* and *Gsp-S1-VII*) and one allele of *Ae. searsii* (*Gsp-S^{s}1-II*).

The N-terminal domain showed five changes, including the presence of the above mentioned InDel in position 36 detected in four alleles of Group II (Fig. 2a). In this same position, the alleles of Group I presented a Gly residue as the standard sequence (GSP-D1); whereas, in the rest of Group II, three alleles showed a Val residue as the GSP-B1 sequence and for the other two a Gly residue. Of the rest of the changes observed in this domain, only one (Ile43 → Val) was found in the variants of Group II, in contrast to Group I where three changes (Glu39 → Gly, Ala44 → Val and Pro45 → Ser) were

detected. Meanwhile, the C-terminal domain was more conservative among overall sequences - all of them had a Trp residue presents (position 164) at the end of this domain as for GSP-B1, while GSP-D1 showed a Leu residue. The presence of a non-conservative change (Tyr162 → Phe) was detected in the protein of the *Gsp-S1-VI* allele (Figs. 2a and Fig.2b).

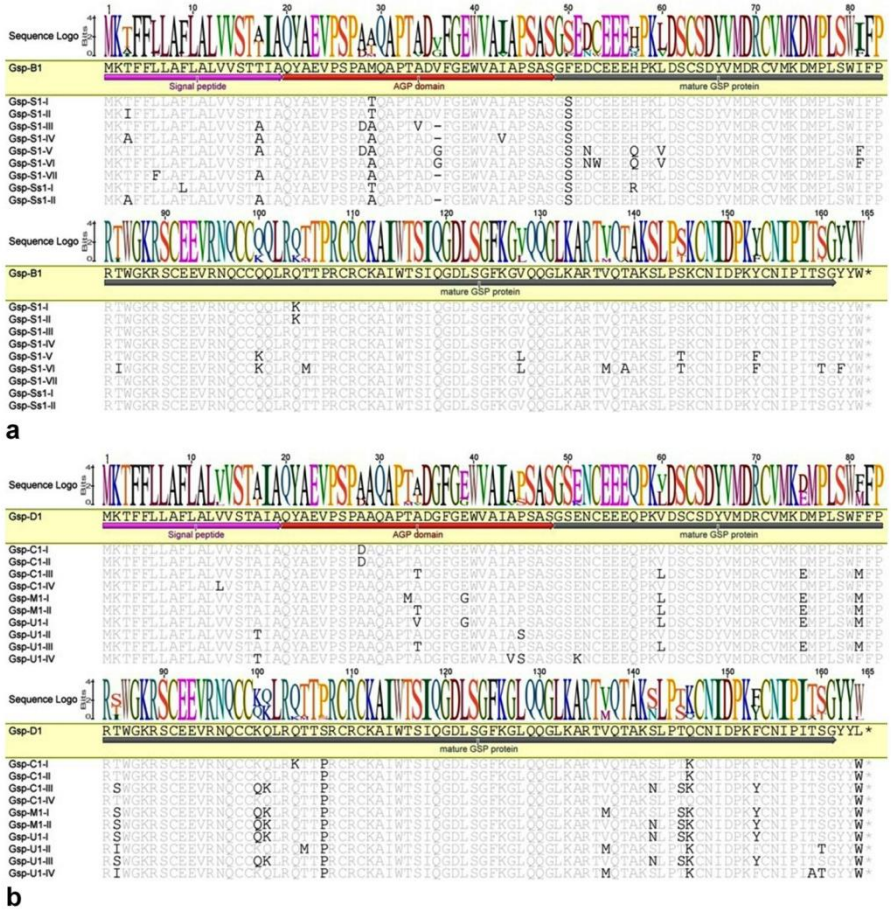


Figure 2. Protein alignment of *Gsp-1* sequences for the species of *Aegilops* with C, M and U genomes (Group I) - **a** -, and with S genome (Group II) - **b** -. For the Sequence Logo, on the y axis (measured in bits), depicts the overall height of the stack indicating the sequence conservation at that position. The height of the symbol indicates the relative frequency of that amino acid at that position.

For the AGP domain, related to the AGP (Glu20-Asp35), seven different motifs were found among the 19 alleles detected in the *Gsp-1* genes - with three being novel (Table 3). AGP1 was the most widespread AGP motif present in each species, except for

Ae. comosa. Inside the AGP peptide there were three important proline residues at positions 25, 27 and 32 frequently separated by 1-3 amino acids, commonly Ser, Ala or Thr. The three proline residues were conserved for all sequences of Groups I and II; however, several other changes were observed. Residues 28, 29, 33 and 34 showed non-conservative substitutions. The *Gsp-C1-I*, *Gsp-C1-II*, *Gsp-S1-III* and *Gsp-S1-V* alleles showed the Ala28 → Asp change. Two different replacements at position 29 (Met29 → Ala/Thr) were detected in alleles of Group II; whereas Thr33 → Met substitution was only found in the *Gsp-M1-I* allele. In position 34, two types of substitution were found: non-conservative (Ala34 → Thr) in the *Gsp-C1-III*, *Gsp-M1-II* and *Gsp-U1-III* alleles, and conservative (Ala34 → Val) in *Gsp-S1-III* and *Gsp-U1-I*.

Table 3. AGP motifs detected in the evaluated materials.

AGP sequence	AGP type	GSP protein
QYAEVPSPAAQAPTAD	AGP1*	GSP-C1-IV, GSP-U1-II, GSP-U1-IV, GSP-S1-IV, GSP-S1-VI, GSP-S1-VII, GSP-S^S1-II
QYAEVPSPATQAPTAD	AGP3*	GSP-S1-I, GSP-S1-II, GSP-S^S1-I
QYAEVPSPDAQAPTAD	AGP7*	GSP-C1-I, GSP-C1-II, GSP-S1-V
QYAEVPSPDAQAPTVD	AGP8*	GSP-S1-III
QYAEVPSPAAQAPMAD	[AGP22]	GSP-M1-I
QYAEVPSPAAQAPTTD	[AGP23]	GSP-C1-III, GSP-M1-II, GSP-U1-III
QYAEVPSPAAQAPTVD	[AGP24]	GSP-U1-I

* according to Wilkinson et al. 2013. The novel AGP motifs appear between brackets.

The two main features in the GSP-1 protein, 10 conserved cysteine residues and one Trp-rich domain, were not conserved through all sequences. Inside the Cys backbone an important mutation was detected: the substitution Cys53 → Trp. This change was only found in the *Gsp-S1-VI* allele and could alter the stability of its structure. For the Trp-rich domain, closed by Cys71-Cys91, the two conserved Trp residues were invariant but had replacements inside this domain that might alter the bond to starch granules; the Asp75 → Glu and Phe81 → Met in Group I alleles (*Gsp-C1-III*, *Gsp-M1-I*, *Gsp-M1-II*, *Gsp-U1-I* and *Gsp-U1-III*); Ile81 → Phe replacement in *Gsp-S1-V* and *Gsp-S1-VI*; Thr85 → Ser change in *Gsp-C1-III*, *Gsp-M1-I*, *Gsp-M1-II*, *Gsp-U1-I* and *Gsp-U1-III*; and Thr85 → Ile substitution in the *Gsp-S1-VI*, *Gsp-U1-II* and *Gsp-U1-IV* alleles.

Other prominent mutations observed in GSP-1 domain were Glu51 → Lys and Thr159 → Ala. Both were observed in the same allele, *Gsp-U1-IV*. The Thr105 → Met

change was observed in the *Gsp-S1-VI* and *Gsp-U1-II* alleles. Finally, the Ser160 → Thr replacement was observed in the *Gsp-S1-VI*, *Gsp-U1-II* and *Gsp-U1-IV* alleles.

Phylogenetic analysis

The complete sequences of the novel alleles obtained from all the *Aegilops* accessions evaluated, together with *Gsp-1* genes sequences present in the GenBank database for *Ae. tauschii*, einkorn and common wheat, were used to construct a phenogram based on the Maximum Composite Likelihood method (Fig. 3). All sequences were arranged in four groups with high bootstrap level. Two of these groups appeared clearly associated with the *Gsp-B1* and *Gsp-D1* genes of wheat.

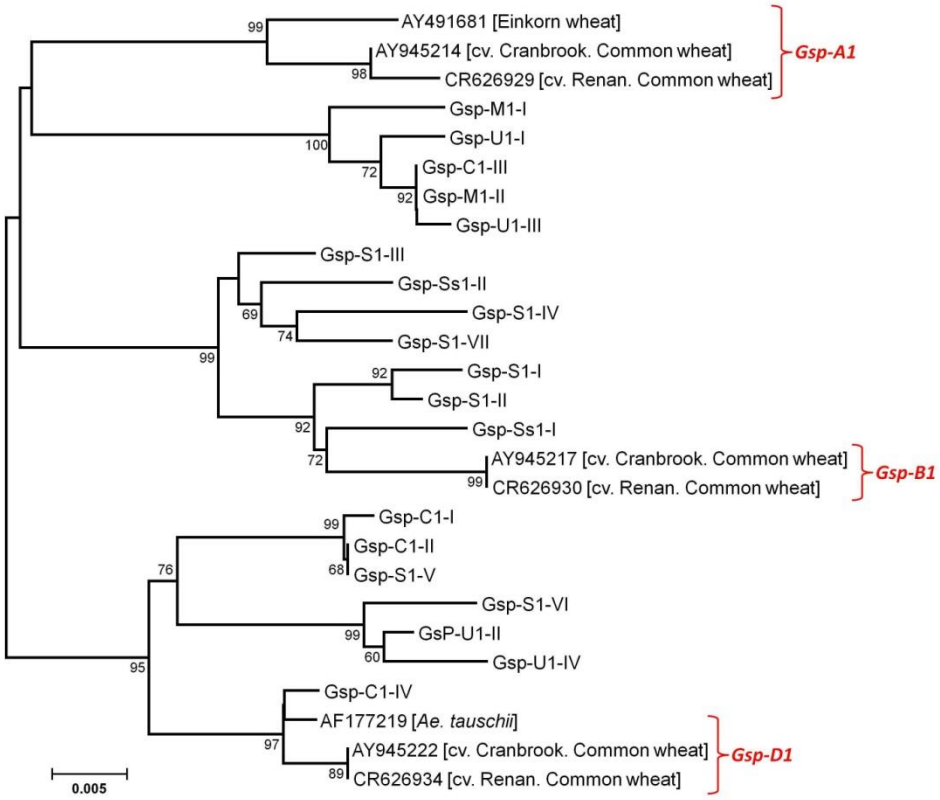


Figure 3. Neighbour-joining tree based on the maximum composite likehood method of *Gsp-1* gene sequences in the *Aegilops* species evaluated, together with other *Gsp-1* published sequences. Numbers in nodes indicate bootstrap estimates from 1,000 replications.

The sequences of the *Sitopsis* species (GroupII) appeared to join with the *Gsp-B1* genes with the exception of the *Gsp-S1-V* and *Gsp-S1-VI* alleles of *Ae. speltoides*, which appeared included with others sequences of the GoupI as two sub-groups into the set associated with the *Gsp-D1* gene: the *Gsp-S1-V* gene with two alleles of *Ae. markgraffi* and the *Gsp-S1-VI* with two ones of *Ae. umbellulata*. The rest of the sequences of GroupI formed one group separated to the other three sets. In addition, the sequences of the genome A showed high homology between them and formed one individual group (Fig. 3).

Discussion

This study examined the polymorphism in *Gsp-1* genes of *Aegilops* species in order to find and characterize novel variability that could be used in modern wheat breeding to develop lines with novel texture. Novel variability was detected, and 11 of the total of 19 different alleles identified were novel. Comparing the current data with previous research on *Pina* and *Pinb* (Cuesta et al. 2013), the variation in *Gsp-1* genes was slightly lower than that of *Pinb*. However, the level of *Gsp-1* polymorphism in the current study did not concur with that obtained by Massa et al. (2004), who found that the *Gsp-D1* genes showed lower polymorphism than *Pinb-D1*; but did concur with that of Gollan et al. (2007), who also found a high level of polymorphism in *Gsp-1* genes from durum wheat. The level of polymorphism reported in the current study was also higher than that shown in wild emmer [*T. turgidum* spp. *dicoccoides* (Körn. ex Asch. & Graebner) Thell.; 2n = 4× = 42, AABB], where *Gsp-A1* polymorphism exceeded that of *Gsp-B1* (Haudry et al. 2007). This high variability in *Gsp-1* could be due to its hypothetical role in defence of plants against bacteria and fungi (Gollan et al. 2007; Philips et al. 2011), as these kinds of genes experience elevated rates of mutations (Chantret et al. 2005). Indirectly, this variation generated might be a source to increase the panel of textures available for wheat breeding.

The GSP-1 protein can be divided into five domains: signal peptide, AGP, N-terminal, GSP-1 and C-terminal. Some researchers have proposed that the AGP peptide (Glu20-Asp35) is related to hardness (Turnbull and Rahman 2002). Thus, mutations in the corresponding part of the sequence could affect grain texture. The most important residues in AGP peptide, three prolines at positions 25, 27 and 32 (Van den Bulck et al. 2002) were conserved in all the accessions. In general, the residues that varied in the AGP sequence in the current study (28, 29, 33 and 34) matched most of those described

by Wilkinson et al. (2013) in an extensive list of species of the *Poaceae* family; however the above mentioned study found more position variations likely due to the greater number of species used. As observed by Wilkinson et al. (2013), AGP1 was the most extensive AGP motif in the *Aegilops* accessions evaluated. Moreover, wheat non-starch polysaccharide as AGP is part of dietary fibre and has been associated with healthy properties of wheat. Fibre helps to delay carbohydrate digestibility decreasing the glycaemic index after a meal, which is relevant to prevention of obesity and diabetes (see Lafiandra et al. 2014 for a review). Other studies have suggested a putative function to the arabinogalactan peptides in metal binding. Therefore, it could be used also as a possible target in breeding for biofortified wheats with enhanced iron and zinc grain concentrations, as Aizat et al. (2011) did with transgenic barley.

The conserved GSP-1 region among the analysed species agrees with the research of Wilkinson et al. (2013), in which most of the residues were highly conserved with only single substitutions. The main features of GSP-1 protein, the cysteine backbone and Trp-rich domain, were not conserved through all sequences. The cysteine backbone, which is shared with PINs, is formed by 10 cysteine residues and is essential for stabilization of the three-dimensional structure and the formation of a lipid-binding hydrophobic cavity (Pauly et al. 2013). The most important mutation was found in the first cysteine residue in the *Gsp-S1-VI* allele, which presented one novel Cys53 $\rightarrow$ Trp substitution. Consequently, this protein is likely to be less stable than the others and therefore could affect grain hardness. In fact, similar changes have been found in PINs resulting in increased hardness (Feiz et al. 2009). To date, mutations at cysteine level in GSP-1 have only been detected by Wilkinson et al. (2013) but not in *Aegilops*.

The Trp-domain has been shown to be the most important region of both PINA and PINB, being essential to preserve the softness texture of grain (Feiz et al. 2009). The hydrophobic Trp residues show an affinity for polar lipids on the surface of starch granules (Pauly et al. 2013). Although, the GSP-1 protein only possesses two Trp residues in this domain, *Gsp-1* genes may also be important in determining grain texture to some extent - especially in durum wheat, which lacks PINs, and in common wheat with identical *Pin* alleles. Moreover, formation of electrostatic bonds between PINs and polar lipids strengthens hydrophobic interactions between the Trp-rich domain and granules starch lipids. In this regard, GSP-1 showed a higher pI than that of PINs, which could enhance these interactions (Phillips et al. 2011). For mature GSP-1 protein, the Trp residues were invariant in all sequences examined in the current study. Inside the Trp-

domain, closed by Cys23-Cys43 (Elmorjani et al. 2013), some replacements were found and changed the properties hydrophobic of this domain. The Thr85 → Ile substitution changed the existing amino acid for a hydrophobic one in *Gsp-S1-VI*, *Gsp-U1-II* and *Gsp-U1-IV* alleles. Another change found in two Group II species (*Gsp-S1-V* and *Gsp-S1-VI*) was the Ile81 → Phe replacement changing the existing residue for a more hydrophobic one. These changes provide greater hydrophobicity and may lead to stronger interactions with lipids on the surface of starch granules, affecting the strength adhesion between starch granules and the surrounding protein matrix (Pauly et al. 2013). Consequently, these replacements could lead to softer texture. However, other changes detected inside the Trp-domain could lead to harder texture, because they might provide less membrane affinity, as for Asp75 → Glu and Phe81 → Met, found in Group I species, since the existing amino acid is replaced by one less hydrophobic.

Therefore, great and novel variability of *Gsp-1* was detected in the present study in different *Aegilops* species. However, it is still not possible to predict the effect of most of the described mutations at the level of protein or grain hardness, since few studies have examined functionality or stability of the GSP-1 protein, and there has been no study of the relationship between polymorphism of these proteins and grain hardness. In recent research (Elmojarni et al. 2013), GSP-1 protein from a soft bread wheat variety was purified for the first time and its lipid interaction checked. The results discarded hydrophobic interactions between GSP-1 protein and lipids; however, only one isoform (*Gsp-D1-b*) was studied. The high homology between PINs and GSP-1 proteins, showing the same principal features (Trp-rich domain and cysteine backbone), suggest that GSP-1 protein could interact with the lipids of starch granules in a similar way to PINs and could plays minor role in grain hardness, although this remains controversial (Morris et al. 2013).

In a previous study carried out with these species (Ortega et al. 2014a), the data confirmed the high similarity between the genomes S and B and the evolutionary proximity among the genomes C, M and U with the genome D of *Ae. tauschii*. In fact, *Ae. umbellulata* (genome U) is included in the same section of the *Aegilops* genus that *Ae. tauschii*. However, in the current study, the *Gsp* genes of the genomes C, M and U showed differences with respect to the genome D, being only the *Gsp-C-IV* allele the sequence associated with this last genome. On the contrary, the sequences of the species of the *Sitopsis* section (*Ae. searsii* and *Ae. speltoides*) were related with the B genome, with exception of two alleles of *Ae. speltoides* that showed similarity with some alleles of

the C and U genomes, which could be related with the previous assumptions suggesting that *Ae. speltoides* is not monophyletic group together with other species of the section *Sitopsis* (Petersen et al. 2006). This suggests that although several authors have suggested that these genes could be useful in phylogenetic studies (see Morris et al. 2013 for a review), their utility could be limited, mainly due to the small size of these genes. This circumstance could magnify the differences between alleles or mask some events of convergent evolution.

In conclusion, high level of polymorphism was detected for these genes in *Aegilops* species, which include differences in amino acid sequence that could lead to changes in hypothetical functionality of the gene product. This should be tested in the genetic pool of modern wheat through conventional breeding or genetic transformation. In parallel, these further studies should determine the true role of these genes in the grain hardness, mainly in durum wheat where their possible effect is not masked by the action of the puroindolines, and valued their effective role as a source of the arabinogalactan peptide (AGP) in wheat non-starch polysaccharide with healthy properties.

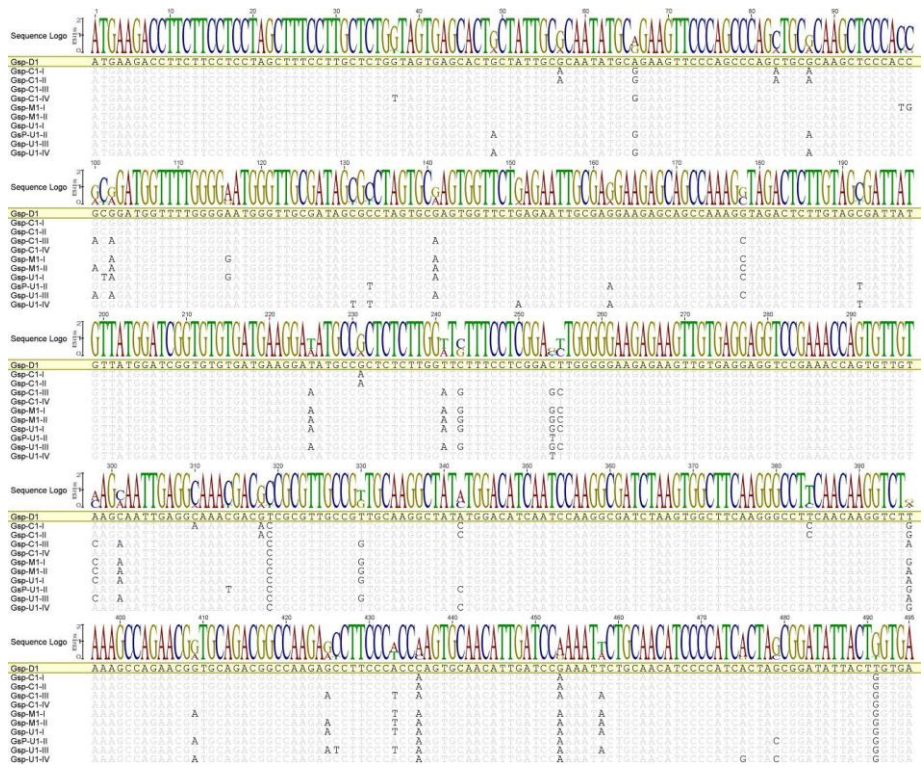
Acknowledgements

This research was supported by grant AGL2010-19643-C02-01 from the Spanish Ministry of Economy and Competitiveness, and the European Regional Development Fund (FEDER) from the European Union. The first author is grateful to the Spanish Ministry of Economy and Competitiveness (FPI programme) and European Social Fund for a predoctoral fellowship. We thank to the National Small Grain Collection (Aberdeen, ID, USA) for supplying the analysed material.

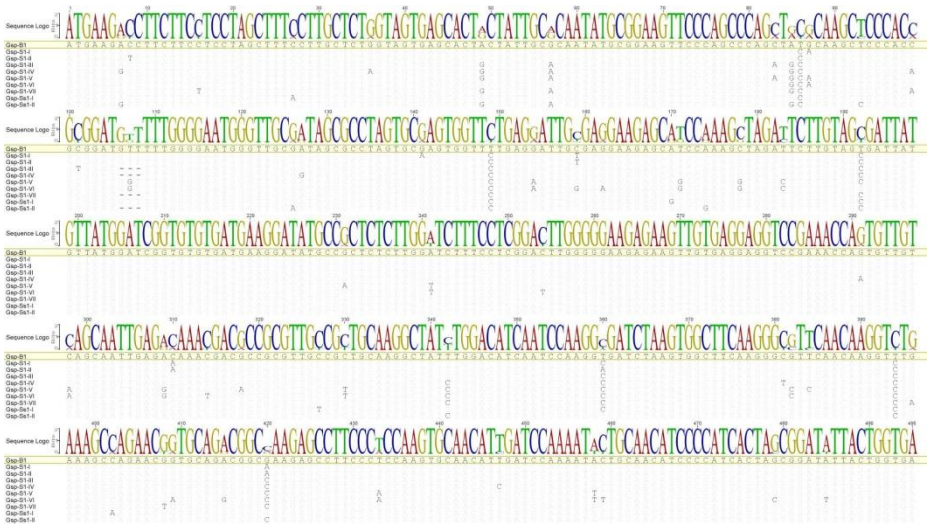
Supplementary material

Supplementary Table 1. *Pina*, *Pinb* and *Gsp-1* alleles in diploids species of the *Aegilops* genus.

Combination	*Pina* allele	*Pinb* alelle	*Gsp-1* alelle	Accession
Ae. comosa				
1	*Pina-M1-I*	*Pinb-M1-I*	*Gsp-M1-I*	PI 551018
2		*Pinb-M1-II*	*Gsp-M1-II*	PI 551053
3		*Pinb-M1-III/II*	*Gsp-M1-II*	PI 551042
4		*Pinb-M1-IV*	*Gsp-M1-I*	PI 551022
5		*Pinb-M1-Null*	*Gsp-M1-II*	PI 551036
6	*Pina-M1-II*	*Pinb-M1-I*	*Gsp-M1-I*	PI 542172
7		*Pinb-M1-III*	*Gsp-M1-I*	PI 551038
8		*Pinb-M1-Null*	*Gsp-M1-I*	PI 551076
9	*Pina-M1-III*	*Pinb-M1-V*	*Gsp-M1-I*	PI 551064
Ae. markgrafii				
1	*Pina-C1-I*	*Pinb-C1-I*	*Gsp-C1-I*	PI 573418
2		*Pinb-C1-II*	*Gsp-C1-I*	PI 254863
3		*Pinb-C1-III*	*Gsp-C1-I*	PI 551129
4		*Pinb-C1-IV*	*Gsp-C1-I/IV*	PI 573413
5	*Pina-C1-II*	*Pinb-C1-I*	*Gsp-C1-I*	PI 564194
6		*Pinb-C1-II*	*Gsp-C1-I/II*	PI 203431
7	*Pina-C1-III*	*Pinb-C1-III*	*Gsp-C1-III*	PI 554237
Ae. searsii				
1	*Pina-Ss1-I*	*Pinb-Ss1-I*	*Gsp-S^s-II*	PI 599152
2		*Pinb-Ss1-II*	*Gsp-S^s-II*	PI 599151
3		*Pinb-Ss-II/III*	*Gsp-S^s-I/II*	PI 599134
4	*Pina-Ss1-II*	*Pinb-Ss-I*	*Gsp-S^s-I*	PI 599174
5		*Pinb-Ss-II*	*Gsp-S^s-II*	PI 599158
Ae. speltoides				
1	*Pina-S1-I*	*Pinb-S1-I*	*Gsp-S1-I*	PI 487233
2		*Pinb-S1-V*	*Gsp-S1-I*	PI 487232
3		*Pinb-S1-VI*	*Gsp-S1-I*	PI 487231
4		*Pinb-S1-VIII*	*Gsp-S1-II*	PI 393493
5	*Pina-S1-II*	*Pinb-S1-I*		PI 486263
6		*Pinb-S1-II*	*Gsp-S1-II*	PI 487236
7		*Pinb-S1-III*	*Gsp-S1-II/IV*	PI 554298
8		*Pinb-S1-VI*	*Gsp-S1-I*	PI 487238
9		*Pinb-S1-VII*	*Gsp-S1-VII*	PI 573448
10		*Pinb-S1-VIII*	*Gsp-S1-II*	PI 219867
11	*Pina-S1-III*	*Pinb-S1-VII*	*Gsp-S1-IV*	PI 554296
12		*Pinb-S1-IV/III*	*Gsp-S1-III*	PI 486262
13		*Pinb-S1-VIII*	*Gsp-S1-I*	PI 170203
14	*Pina-S1-IV*	*Pinb-S1-IX*	*Gsp-S1-V/VI*	PI 554304
Ae. umbellulata				
1	*Pina-U1-I*	*Pinb-U1-I*	*Gsp-U1-II*	PI 298906
2		*Pinb-U1-III*	*Gsp-U1-I/II*	Ciae 66
3		*Pinb-U1-Null*	*Gsp-U1-II*	PI 542365
4	*Pina-U1-II*	*Pinb-U1-III*	*Gsp-U1-III*	PI 554417
5	*Pina-U1-III*	*Pinb-U1-I*	*Gsp-U1-II*	PI 542372
6		*Pinb-U1-II*	*Gsp-U1-II*	PI 542377
7	*Pina-U1-IV*	*Pinb-U1-I*	*Gsp-U1-IV*	PI 560556



Supplementary Figure 1. Nucleotide alignment of *Gsp-1* sequences for the species of *Aegilops* with C, M and U genomes (Group I).



Supplementary Figure 2. Nucleotide alignment of *Gsp-1* sequences for the species of *Aegilops* with S genomes (Group II).

CAPÍTULO III

CARACTERIZACIÓN MOLECULAR DE NUEVOS GENES LMW-m Y LMW-s EN CUATRO ESPECIES DEL GÉNERO *Aegilops* (SECCIÓN *Sitopsis*) Y COMPARACIÓN CON LOS DEL LOCUS *Glu-B3* EN TRIGO COMÚN

Enviado como:

S. Cuesta, C. Guzmán, J.B. Alvarez (2015) Molecular characterization of novel LMW-m and -s genes from four *Aegilops* species (*Sitopsis* section) and comparison with those from the *Glu-B3* locus of common wheat. *Molecular Breeding* (under review).

Resumen

Las subunidades de bajo peso molecular de glutenina (LMWGs) son un componente del gluten que juega un papel en la determinación de las propiedades viscoelásticas de la masa de harina de trigo. Las especies de *Aegilops* han demostrado ser una importante fuente de variación en importantes características para la mejora del trigo. Sin embargo, se conoce muy poco sobre los genes de LMWG en especies de la sección *Sitopsis*, la cual está estrechamente relacionada con el genoma B del trigo común. Diez entradas de la sección *Sitopsis* fueron evaluadas para la variabilidad de los genes de LMWG y 20 genes nuevos fueron obtenidos: nueve fueron genes LMW-m y 11 fueron LMW-s. Solo dos fueron pseudogenes, correspondiendo a un gen de LMW-m y uno de LMW-s. Se detectaron seis grupos de genes: tres tanto para genes de LMW-m como para genes de LMW-s. Todos los grupos de los genes de LMW-s y uno de los genes de LMW-m (*pGluU*) detectados no estuvieron relacionados a los genes del genoma B del trigo común, mientras que el resto de genes si lo estuvieron. El polimorfismo de un solo nucleótido y los InDels detectados en las variantes activas comparadas con las de trigo común podrían afectar a la estructura de la proteína e influir en la calidad de la masa. El análisis de los epítopos reactivos para la enfermedad celíaca reveló que las subunidades de LMW-s carecían de toxicidad, así como las subunidades de LMW-m del grupo *pGluU*, mientras que las otras subunidades de LMW-m fueron menos tóxicas que las de trigo común.

Palabras clave: *Aegilops* sp., enfermedad celíaca, calidad del gluten, genes de LMWG, *Sitopsis*.

Abstract

Low molecular-weight glutenin subunits (LMWGs) are a component of the gluten network that play a key role in determining the viscoelastic properties of wheat dough. *Aegilops* species have been shown to be an important source of variation in valuable traits for wheat breeding. However, very little is known about LMWG genes in *Sitopsis* species, which are closely related to the B genome of common wheat. Ten accessions of *Sitopsis* species were evaluated for variability of LMWG genes and 20 novel genes were obtained: nine were LMW-m and 11 were LMW-s genes. Only two were pseudogenes, corresponding to one LMW-m and one LMW-s gene. Six groups of genes were detected: three for each of the LMW-m and LMW-s genes. All groups of LMW-s genes and one of LMW-m genes (*pGluU*) detected were not related to B-genome genes from common wheat, whereas the remaining genes were. The single nucleotide polymorphisms and InDels detected in active variants compared to those from common wheat could affect structure protein and influence dough quality. The analysis of reactive epitopes for celiac disease revealed that LMW-s subunits lacked toxicity, as well as the *pGluU* LMW-m subunits, whereas the other LMW-m subunits were less toxic than that from common wheat.

Key words: *Aegilops* sp., celiac disease, gluten quality, LMWGs genes, *Sitopsis*.

Introduction

Wheat gluten, the network responsible for wheat flour forming dough with viscoelastic properties suitable for the production of diverse foods (see Wrigley et al. 2006 for reviews), is mainly formed by interaction of two components: glutenins and gliadins. Glutenins are divided in two types of subunits: high molecular-weight and low molecular-weight (LMWGs). These latter subunits represent 60% of the glutenin fraction and are classified into three main types by the first amino-acid residue of the mature protein: LMW-i (isoleucine - Ile), LMW-m (methionine - Met) and LMW-s (serine - Ser). Their role in pasta-making quality is well established (see Rasheed et al. 2014 for reviews), being associated with gluten strength. Their primary structure consists of an N-terminal domain, a central repetitive domain and finally a C-terminal domain with cysteine (Cys) residues involved in three inter- and two intra-molecular disulphide bonds (D'Ovidio and Masci 2004). The repetitive domain is mainly responsible for the length of LMWGs, whereas the C-terminal domain contains most of the Cys residues. Alterations as amino-acid substitutions and InDels may generate changes with an effect on dough quality (Masci et al. 1998; Tanaka et al. 2005; Chen et al. 2011). In addition, different studies have indicated the presence of reactive epitopes inside LMWGs that are related to celiac disease (Rasheed et al. 2014; Cuesta et al. 2015).

Nevertheless, genetic study of LMWGs has been complicated due to the great number of subunits encoded by this multigene family. In common wheat (*Triticum aestivum* L. ssp. *aestivum*; 2n = 6× = 42, DDAABB), LMWG genes are encoded at *Glu-3* loci on the short arms of chromosomes 1A, 1B and 1D (Pogna et al. 1990; Singh and Shepherd 1988), with each locus formed by several LMWG genes whose exact number remains unknown (Rasheed et al. 2014). However, Zhang et al. (2013) identified more than 15 LMWG genes grouped in two clusters (m_{AD} and i_A) for the *Glu-A3* locus, two others (m_{BD} and s_{BD}) for the *Glu-B3* locus, and one cluster (m_D-2-s_{BD}) together with some unlinked genes (m_{BD}, m_{AD} and m_D-1) for the *Glu-D3* locus. Each cluster is associated with one LMWG type. In the *Glu-B3* locus analysed in the current study, the LMW-s subunits were synthesised by five genes (*B3-544*, *B3-578*, *B3-621*, *B3-688* and *B3-813*) grouped in the s_{BD} cluster, while the m_{DB} cluster was formed by two LMW-m genes (*B3-530* and *B3-548*).

Analysis of the nucleotide sequences of variants of these genes has permitted the establishing of relationships with previous studies that used other denominations or classifications. The *B3-544* gene corresponds to previously identified *GluB3-1* (Wang et al. 2009) and *0154F22-s* (Huang and Cloutier 2008), and *B3-621* corresponds to *GluB3-2* (Wang et al. 2009). *B3-688* was previously named *GluB3-3* by Wang et al. (2009) and is related to Y17845 sequences associated with good quality and *B3-2*, *B3-3* and Group 3 type-II genes (Dong et al. 2010; Ikeda et al. 2002; Masci et al. 1998). Of LMW-m genes, the *B3-530* gene was previously characterised as *GluB3-4* by Wang et al. (2009), and is associated with *1557N24-m*, *B3-1* and the Group 2 type-I genes (Dong et al. 2010; Huang and Cloutier 2008; Ikeda et al. 2002).

These genes have been also studied in some wheat relatives, mainly in such donor species of the wheat genomes as *T. urartu* Thum. ex Gandil for the A genome (Cuesta et al. 2015; Long et al. 2008; Lou et al. 2015) or *Aegilops tauschii* Coss. for the D genome (Johal et al. 2004; Pei et al. 2007; Zhao et al. 2008). However, due to the uncertain and controversial origin of the B genome (Huang et al. 2002), the putative donors of this genome have been scarcely studied (Jiang et al. 2008; Huang et al. 2010; Li et al. 2010; Lin-Hai et al. 2010; Wang et al. 2011a). Although *Ae. speltoides* Tausch is now considered the most likely progenitor of the B genome (Petersen et al. 2006), four other *Aegilops* species from the *Sitopsis* section [*Ae. bicornis* (Forssk.) Jaub. & Spach., *Ae. longissima* Schweinf. & Muschl., *Ae. searsii* Feldman & Kislev ex K. Hammer and *Ae. sharonensis* Eig.] have also been considered as possible donors of this genome (see Haider et al. 2013 for reviews). Additionally, in consonance with previous studies that implied alternatives to the five *Aegilops* species as its origin, some studies have suggested the possibility of a polyphyletic origin of the B genome (Sarkar and Stebbins 1956).

The utility of *Aegilops* species in wheat improvement has been indicated (see Schneider et al. 2008 for reviews); some of these species have shown great and novel variability in such valuable traits as disease and pest resistance, and stress and salt tolerance that can be used in wheat breeding. Concerning flour quality, study of LMWG genes has revealed that the *Aegilops* genus possesses a large variation that could play an important role in improving the dough properties of wheat (Zhao et al. 2008; Chen et al. 2010; Wang et al. 2011b). However, few LMWG genes from section *Sitopsis* have been reported, and studies on their characterisation are also scarce (Huang et al. 2010; Jiang et al. 2008; Li et al. 2010; Lin-Hai et al. 2010; Wang et al. 2011a).

The aim of the current study was the molecular characterisation of LMWG genes in 10 *Aegilops* accessions from section *Sitopsis*, together with the comparison and analysis of the relationships between the *Glu-B3* genes of common wheat.

Materials and methods

Plant vegetal

Three accessions of *Ae. longissima* and *Ae. sharonenesis*, and two of *Ae. searsii* and *Ae. speltoides* were used in this study (Table 1). These accessions were obtained from the National Small Grains Collection (Aberdeen, Idaho, USA).

DNA analysis: extraction, amplification and sequencing

Genomic DNA was isolated from young leaves of a single plant per accession according to the method of CTAB (Stacey and Isaac 1994). The complete coding region of the LMWGs genes, together with 104 bp of the 3'-UTR, was amplified with the primers designed by Ma et al. (2006): 5'-ATGAAGACCTTCCTCGTCTTT-3' and Zhang et al. (2011): 5'-TCACACATGACGTTGTGTGAC-3'. PCR amplification of genomic DNA was performed in a volume total of 20 μl containing 50 ng of genomic DNA, 0.3 μM of each primer, 0.4 mM of dNTPs, 1 mM or 2mM of $MgCl_2$, 1× of reaction buffer and 1 U of *Taq* DNA polymerase (Promega, Madison, WI, USA). The amplification was carried out with a first step of initial denaturation at 94 °C for 3 min followed by 35 cycles of 30 s of denaturation at 94 °C, a step of annealing of 30 s at 58 °C, then 1 min. of extension at 72 °C. To finish, an extension final step at 72 °C for 10 min was performed. The PCR products (amplicons) were separated by electrophoresis on a 1.2% agarose gel, stained with ethidium bromide and visualised under UV light.

These amplicons were purified using Sureclean (Bioline), ligated to pGEM-T (Promega, Madison, WI, USA) and used to transform *Escherichia coli* JM109 competent cells. At least three different inserts were sequenced. The novel sequences are available from GenBank database.

Data analysis

The sequences obtained in the current study were analysed and compared with the sequences available in GenBank database by BLAST analysis using the Geneious Pro ver. 5.0.3 software (Biomatters Ltd.). Phylogenetic tree was constructed with MEGA5 software (Tamura et al. 2011) using LMWGs sequences obtained together with LMWGs

genes previously identified in *Glu-B3* locus from common wheat by Zhang et al. (2013) and other authors (Ikeda et al. 2002; Huang and Cloutier 2008; Wang et al. 2009; Dong et al. 2010), also as the LMW-s gene isolated from common wheat cv. Yecora Rojo (NCBI: Y17845; Masci et al. 1998). LMW-s and LMW-m sequences isolated from *Sitopsis* species by Jiang et al. (2008), Li et al. (2010), Huang et al. (2010) and Lin-Hai et al. (2010) and five ones from GenBank database were also used to compare. Neighbour-joining cluster with all sequences analysed was generated using the Maximun Composite likelihood method (Tamura et al. 2004) and one bootstrap consensus from 1000 replicates was used (Felsenstein 1985).

Results

Variation of LMWGs genes

Twenty sequences of LMWG genes, two for each accession evaluated, were obtained (Table 1). The BLAST analysis carried out indicated that any of these sequences had been previously identified or catalogued in GenBank. Two of these sequences were pseudogenes: KT156622, detected in *Ae. searsii* (PI 599174 accession), had an insertion of one adenine in the C-I domain that generated one frame-shift mutation; and KT156624 (*Ae. sharonensis*, PI 584363 accession) had two in-frame stop codons, one in the repetitive domain and the other in the C-II domain. According to the first amino acid of deduced mature protein, the former was catalogued as LMW-m, while the latter was LMW-s. However, KT156624 had notable differences compared to the other LMW-s evaluated, with three large deletions (24, 42 and 66 bp) in the repetitive domain, together with two insertions of 18 and 27 bp inside the C-II and C-III domains, respectively.

The remaining sequences were identified as genes: eight were classified as LMW-m with a coding region of range of 897–1113 bp; and 10 as LMW-s with sizes of 996–1086 bp (Table 1). In all cases, the main differences in size were a consequence of InDels in the repetitive domain.

Table 1. LMW-m and LMW-s sequences isolated from *Aegilops* in this study.

Accession	NCBI ID		Region coding size (bp)	LMWG type
Ae. longissima Schweinf. & Muschl.				
PI 604117	KT156614	Gene	1017	LMW-s
	KT156615	Gene	1023	LMW-m
PI 604120	KT156616	Gene	1059	LMW-s
	KT156617	Gene	1041	LMW-s
PI 604121	KT156618	Gene	1098	LMW-m
	KT156619	Gene	1086	LMW-s
Ae. searsii Feldman & Kislev ex K. Hammer				
PI 599151	KT156620	Gene	897	LMW-m
	KT156621	Gene	1059	LMW-s
PI 599174	KT156622	Pseudogene	900	LMW-m
	KT156623	Gene	1059	LMW-s
Ae. sharonensis Eig.				
PI 584363	KT156624	Pseudogene	996	LMW-s
	KT156625	Gene	1059	LMW-s
PI 584378	KT156626	Gene	1074	LMW-m
	KT156627	Gene	1041	LMW-s
PI 584416	KT156628	Gene	1059	LMW-s
	KT156629	Gene	1017	LMW-s
Ae. speltoides Tausch				
PI 487236	KT156630	Gene	900	LMW-m
	KT156631	Gene	1113	LMW-m
PI 554298	KT156632	Gene	1095	LMW-m
	KT156633	Gene	1092	LMW-m

The relationships of these sequences with other LMWGs detected for the *Glu-B3* locus in common wheat were evaluated by the construction of a phenogram based on the Maximum Composite Likelihood method (Fig. 1). In this analysis, other LMWG sequences from *Sitopsis* species obtained by several authors were also included. All sequences were grouped in three main clusters: one included all LMW-s genes, while the LMW-m genes were in two clearly separated clusters (Fig. 1). The variants detected by Zhang et al. (2013) were used as references in the analysis of each of these major clusters and their subdivisions or sets.

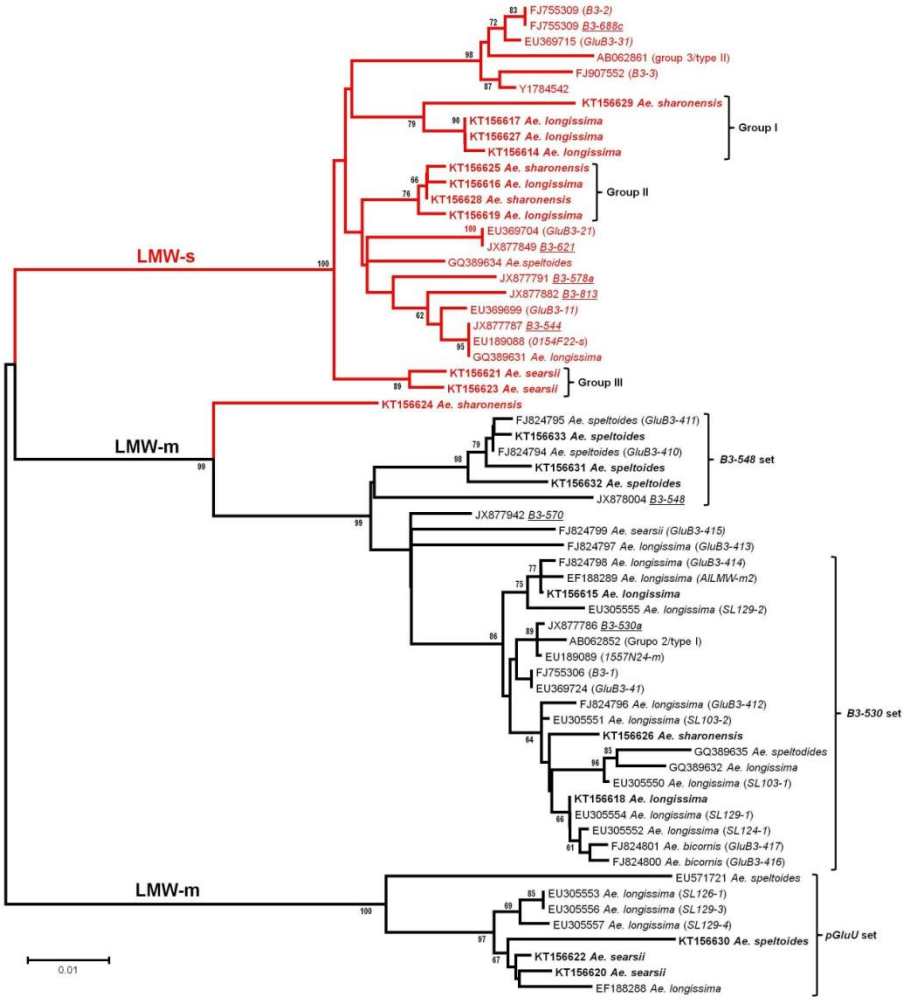


Figure 1. Neibour-joining tree based on the maximum composite likehood method of LMW-s (red) and LMW-m (black) genes sequences obtained in the current study (bold), together with the previous sequences described in common wheat and *Aegilops* species (*Sitopsis* section). Numbers in nodes indicate boot-strap estimates from 1000 replications. The wheat sequences used as references appear underlines.

The LMW-s cluster was formed by four sets: two appeared associated with variants detected in common wheat. One set was only formed by sequences of wheat and was associated with the *B3-688* gene. In another set (Group-II), the sequences KT156616 and KT156619 from *Ae. longissima* and KT156625 and KT156628 from *Ae. sharonensis* were clustered together with some of the variants described in common wheat (*B3-544*, *B3-578*, *B3-621* and *B3-813*), although they formed a well-differentiated subgroup within

it (Fig. 1). The other four LMW-s detected in these species (KT156614 and KT156617 from *Ae. longissima*; KT156627 and KT156629 from *Ae. sharonensis*) formed an independent set (Group-I). The fourth set (Group-III) was formed by the sequences KT156621 and KT156623 from *Ae. searsii*. One special case was the sequence KT156624 (pseudogene) detected in *Ae. sharonensis* that appeared enclosed within one of the LMW-m clusters but was clearly separated from the other LMW-m sequences (Fig. 1).

Of the LMW-m genes, one of the two clusters included the *B3-530* and *B3-570* genes from common wheat, together with the pseudogene *B3-548* (Fig. 1). Three alleles sequenced in this study from *Ae. longissima* (K156615 and KT156618) and *Ae. sharonensis* (KT156626) were associated with the *B3-530* gene. Within this set were also included other sequences previously obtained in *Ae. bicornis* (Huang et al. 2010; Lin-Hai et al. 2010), *Ae. longissima* (Jiang et al. 2008) and *Ae. speltoides* (Lin-Hai et al. 2010). In contrast, the three active sequences obtained from *Ae. speltoides* (KT156631, KT156632 and KT156633) in the current study were associated with the *B3-548* pseudogene. No sequences were related to the *B3-570* gene.

The other LMW-m cluster showed a clear separation from the rest of the sequences, both LMW-m and LMW-s, and comprised two alleles (KT156620 and KT156622) from *Ae. searsii* and one (KT156630) from *Ae. speltoides.* Among the sequences obtained previously (Huang et al. 2010; Li et al. 2010), four sequences from *Ae. longissima* (SL126-1, SL129-3, SL129-4 and EF188288) and one from *Ae. speltoides* (EU571721) were grouped in this cluster, which was named the pGluU set (Fig. 1).

Characterisation of the deduced amino-acid sequences

The deduced mature proteins of the sequences obtained in the current study showed a size range of 278–350 residues for LMW-m and 318–341 for LMW-s (Table 2), and were compared with the proteins from *B3-530a* and *B3-688c* genes in common wheat (Figs. 2 and 3). For overall sequences, no extra or missing Cys residues were detected. However, sequences of the *pGluU* set showed a change in the position of the first Cys residue located in the N-terminal domain (position 5). For the other sequences, this Cys residue was located in the repetitive domain: position 45 for LMW-m or position 46 for LMW-s. The contents of glutamine (Gln) and proline (Pro) residues were similar in all sequences - with median values of 34.9 and 15.3%, respectively, with the exception

of the two sequences included in the *pGluU* set, which had lower values (Gln: 30.6%; Pro: 13.3%).

Table 2 Characteristics of the deduced mature protein of the LMW-m and LMW-s sequences from *Aegilops*.

Set	N-terminal domain	NCBI Id	Mature protein (aa)	Content Gln (%)	Content Pro (%)	Repetitive (aa)	Motifs N	Motifs Size
LMW-m genes								
B3-530[a]	METSHIP/LSLEKS/PL	KT156615	320	34.4	14.7	123	18	3-9
		KT156618	345	34.5	15.9	148	21	3-9
		KT156626	337	34.1	15.4	140	20	3-9
B3-548[a]	METSHL/IPGLENPS	KT156631	350	34.9	16.0	153	22	3-9
		KT156632	344	33.7	15.4	147	21	3-9
		KT156633	343	34.4	16.0	146	21	3-9
pGluU[b]	METSCIPSLERPW	KT156620	278	30.6	12.9	84	13	3-11
		KT156630	279	30.5	13.6	84	13	3-11
LMW-s genes								
Group I	MENSHIPGLERPS	KT156614	318	35.2	14.8	133	19	3-9
		KT156627	326	35.6	15.0	141	20	3-9
		KT156629	318	34.3	14.8	133	19	3-9
	MENSHILGLERPS	KT156617	326	35.6	14.7	141	20	3-9
Group II	MENSHIPGLERPS	KT156616	332	35.5	14.8	149	21	3-9
		KT156619	341	36.1	15.5	155	22	3-9
		KT156625	332	35.2	15.7	147	21	3-9
		KT156628	332	35.8	15.1	149	21	3-9
Group III	MESSHILGLERPS	KT156621	332	34.6	15.4	147	21	3-9
		KT156623	332	34.6	15.4	147	21	3-9

[a] according to Zhang et al. (2013); [b] according to Huang et al. (2010).

The abovementioned groups showed some changes within the N-terminal and could be classified according to the sequence of this domain (Table 2). For LMW-m, the *B3-548* set contained the METSHL/IPGLENPS domain, the *B3-530* set contained METSHIP/LSLEKS/PL and the *pGluU* set had METSCIPSLERPW (Fig. 2). For LMW-s (Fig. 3), Group-I and -II shared the same N-terminal that the reference sequence (MENSHIPGLERPS), except for KT156617, which showed one replacement (MENSHILGLERPS). Group-III showed one remarkable substitution between the two changes detected: the asparagine residue in position 3 specific to LMW-s subunits was replaced by a Ser residue (MESSHILGLERPS).

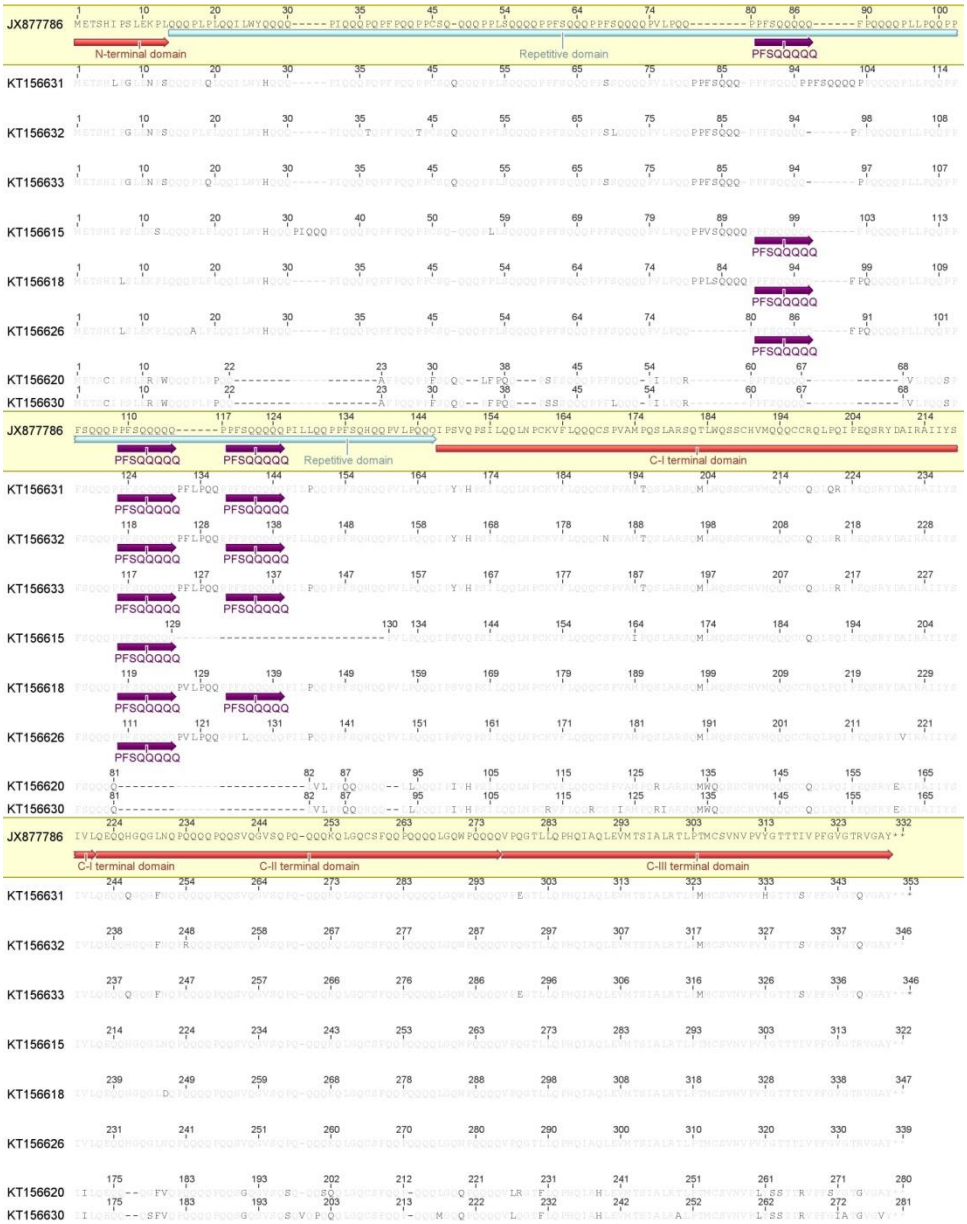


Figure 2. Alignment of the deduced amino acid sequences of LMW-m genes from *Sitopsis* species with respect to the *B3-530a* allele (JX877786) identified by Zhang et al. (2013), and identification of Glt-17 (PFSQQQQQ) epitope.

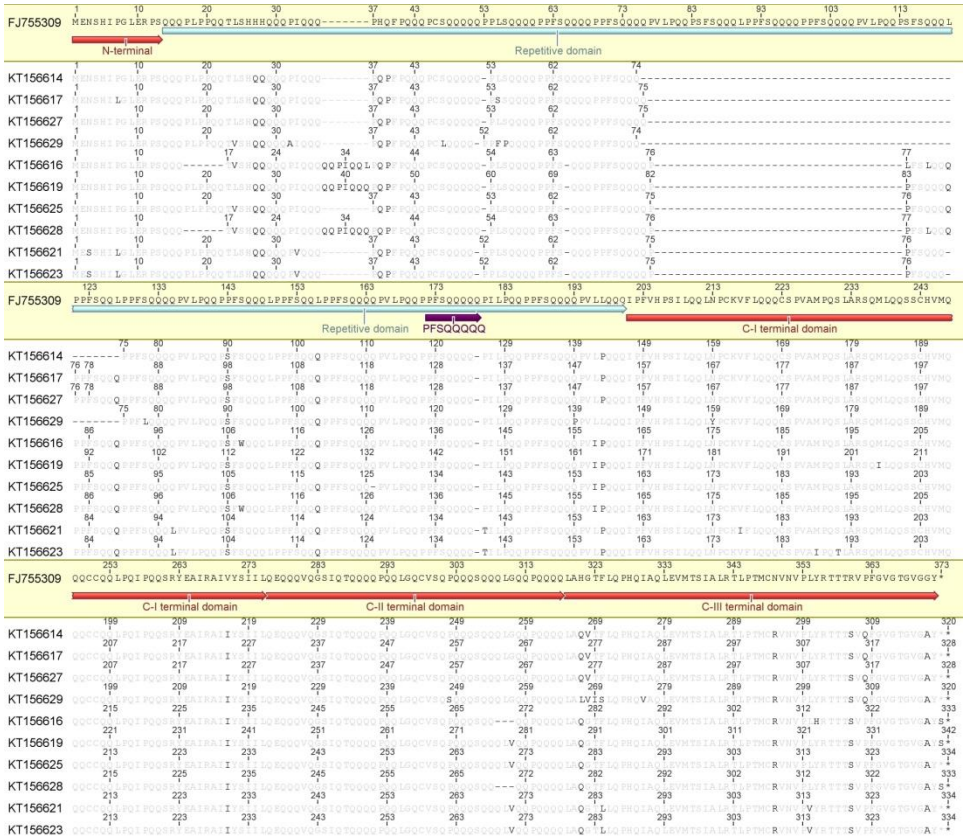


Figure 3. Alignment of the deduced amino acid sequences of LMW-s genes from *Sitopsis* species with respect to the *B3-688c* allele (FJ755309) identified by Zhang et al. (2013), and identification of Glt-17 (PFSQQQQQ) epitope.

The observed differences in protein size were mainly related to the size of the repetitive domain: 84-153 residues for LMW-m and 133-155 in LMW-s. In the LMW-m, this domain was formed by 13–22 repeat motifs of 3–11 residues; in contrast LMW-s had a range of 19-21 repeat motifs formed by 3-9 residues (Table 2). Within this domain, the presence of toxic epitopes considered as T-cells stimulatory for celiac disease (Valder et al. 2002) was determined, although only the Glt-17 epitope (PFSQQQQQ) was found in LMW-m sequences. This epitope was detected up to three times in the KT156618 sequence of LMW-m, but due to several InDels and one substitution in LMW-m was found only twice in the *B3-548* set and the other sequences of the *B3-530* set (Fig. 2). In contrast, the sequences of the *pGluU* set did not show this reactive motif. For overall LMW-s groups, no epitopes were detected due to the presence of one InDel (Fig. 3).

In the repetitive domain was found the most important InDels responsible for their size differences (Figs. 2 and 3). The LMW-m alleles had 1-8 residue insertions in the *B3-530* and *B3-548* sets, together with the deletion of Glu88 residue in the KT156632 and KT156633 sequences. Furthermore, the KT156615 sequence had one large deletion of 29 residues. The *pGluU* set showed clear differences compared to the other LMW-m sequences. Several short unique deletions in the range of 1-8 residues and three large deletions of 13-26 residues were detected in both the repetitive and the C-II domains. Furthermore, within the C-II domain of KT456630 was an insertion of one valine residue between positions 249 and 250 (Fig. 2). Overall LMW-s alleles showed one extended deletion in the repetitive domain, of 44-52 residues in Group-I and 36 residues in Group-II and -III, together with the deletion of Glu179 residue (Fig. 3). Extra deletions of Gln or Pro residues were detected among the groups. Furthermore, within Group-II, the KT156616 and KT156628 had an additional deletion (PLPPQQ), between residues 17 and 22 of the repetitive domain and another (QLG) between residues 309 and 311 in the C-II domain. Three sequences (KT156616, KT156619 and KT156628) had one unique insertion (QQPIQQQ/L) between residues 36 and 37.

The other changes detected in the repetitive and C-terminal domains were substitutions of amino acid residues. Among these changes, those affecting Pro and Gln residues were the most important due to being involved in the maintenance of the LMWG structure. In the *pGluU* set of LMW-m sequences, a great number of positions showed variation (59), with 27 of them affecting Gln and Pro residues, whereas the *B3-548* and *B3-530* sets had less variation (24 and 13, respectively), and only 14 and 7 positions involved changes in Gln and Pro residues. For LMW-s alleles, the number of positions changed was 27 for Group-I and 23 for Group-II and -III, of which 14 and 12 positions, respectively, were involved in changes to Gln and Pro residues.

Discussion

In the search for gene variants from wheat relatives that could be useful as sources of variation for wheat breeding, is important to know the relationships of these relatives with each one of the hexaploid wheat genomes. In the case of the B genome, the main problem is its uncertain origin. So, although *Ae. speltoides* has been suggested as the closest relative to the B genome of polyploid wheat (Petersen et al. 2006), numerous contradictory studies exist and there are alternative explanations for the origin of the B genome (Haider et al. 2013). It has been suggested that the B-genome donor may be

extinct or not yet collected; another theory proposed a polyphyletic origin and that it is the recombination of different donor *Sitopsis* species (Sarkar and Stebbins 1956). In agreement with this hypothesis, our studies with the *Wx* genes in spelt wheat (Guzmán et al. 2012), showed that the *Wx-B1* gene could have at least two different origins, while the orthologous *Wx-A1* and *Wx-D1* were monophyletic. A later study suggested that these two origins could be associated with *Ae. searsii* and *Ae. speltoides*, respectively (Ortega et al. 2014a). For wheat breeding, this multiplicity of origins could be more an advantage than a problem because this would significantly expand the sources of variation.

In the current study of LMWGs, four *Aegilops* species (sect. *Sitopsis*) were analysed for these genes to evaluate them as alternative sources of variation for increasing the genetic background of modern wheat. In this respect, all comparisons of the LMWG sequences obtained from these *Aegilops* species were carried out with the gene components of the *Glu-B3* locus described in common wheat by Zhang et al. (2013) and grouped within the m_{BD} (LMW-m) and s_{BD} (LMW-s) clusters. Three separate groups were found inside the nine LMW-m sequences detected in the current study. The two sequences obtained from *Ae. longissima* and one from *Ae. sharonensis* were associated with the *B3-530* gene, which is the main active gene of the m_{BD} cluster from common wheat (Huang and Cloutier 2008; Wang et al. 2009; Dong et al. 2010; ; Zhang et al. 2013). In this group were also included the LMWG sequences from *Ae. longissima* obtained by Huang et al. (2010) using *pGluB* (B genome specific) primers. In contrast, three of the four sequences from *Ae. speltoides* were associated with *B3-548*, a pseudogene; although, in this case, the overall sequences were active. The remaining LMW-m sequences (two from *Ae. searsii* and one from *Ae. speltoides*) were included in an additional cluster that coincided with the LMW-m sequences amplified with the LMWG universal primers (*pGluU*) designed by Huang et al. (2010), which showed association with the genes grouped in the m_{AD} cluster from *Glu-A3* and *Glu-D3* loci by Zhang et al. (2013). However, the LMW-s sequences evaluated here showed no association with the genes (*B3-621* or *B3-688*) contained in the two haplotypes described for the s_{BD} cluster in common wheat by Zhang et al. (2013).

Independently of the importance of this variation to clarify the origin and evolutionary events that generated the actual B genome in wheat, the main characteristics to be evaluated for determining any interest in these novel LMWGs variants for wheat breeding is the internal structure of the mature protein because this could affect their functionality (Ma et al. 2006). Several aspects of these subunits were evaluated due to

effects on flour quality previously indicated by other authors (Masci et al. 1998, 2000; Ikeda et al. 2002; D'Ovidio and Masci 2004; Tanaka et al. 2005): number and distribution of Cys residues, size and structure of the repetitive domain, or the content of Pro and Gln residues.

The LMWGs have a highly conserved backbone of eight Cys residues, which are involved in the formation of intra- and inter-molecular disulphide bonds; and any modification in the number or position of these residues could affect gluten strength (Masci et al. 1998). The distribution of Cys residues was used by Ikeda et al. (2002) to classify the LMWGs into six types, although D'Ovidio and Masci (2004) later simplified this classification to three main types: those with the first Cys residue in the N-terminal domain, those with the first Cys in the repetitive domain and those with all Cys in the C-terminal domain. However, the impact on dough quality should be demonstrated. Ikeda et al. (2002) suggested that the third type could lead to functional differences compared to the other two types. Sixteen of the sequences obtained in the present study corresponded with subunits of the second type; while the two sequences of the *pGluU* set were of the first type due to the presence of the first Cys residue in position 5 of the N-terminal domain. These latter sequences also showed lower percentages of Gln and Pro content (30.6 and 13.3%, respectively) than the other sequences, with corresponding median values of 34.9 and 15.3%. Furthermore, both sequences had the smallest repetitive domains, whereas in the LMW-s subunits, no reactive epitope for celiac disease was found, while the rest of the LMW-m subunits had these epitopes. Nevertheless, the number of epitopes found for LMW-m subunits was lower than that for reference sequences of common wheat and that found for the LMW-i subunits in a previous study of *T. urartu* Thum ex. Gandil. (Cuesta et al. 2015), but slightly higher than that detected in LMW-m from a diploid *Triticum* species (Cuesta et al. unpublished results).

Because other authors have suggested that the changes in the percentage of both amino acids and the variation in the size of the repetitive domain could influence gluten strength (Masci et al. 1998, 2000; Tanaka et al. 2005), the effect of these novel alleles should be evaluated in future studies when these genes have been introgressed into wheat. Additionally, due to the lack of reactive epitopes, the LMW-m of the *pGluU* set, together with the LMW-s, could have an advantage in a strategy for developing wheat cultivars with new flour properties and that are suitable for people with celiac disease.

As mentioned above, the main differentiation between LMW-m and LMW-s genes is based in the first amino acid of the N-terminal domain in the mature protein (Met

and Ser, respectively). However, paradoxically, the N-terminal domain of the LMW-s of Group-III did not begin with Ser. The typical sequences of this N-terminal domain in LMW-s are MENSHIPGLERPS- and IENSHIPGLEKPS-; being other two characteristics the presence of a specific peptide TLSH (with some variants), together with a unique position of seven Cys residues at C-terminal II (Ikeda et al. 2002; Luo et al. 2015). The sequences obtained from *Aegilops* species in the current study showed similar characteristics. Recent studies have shown that the mature LMW-s proteins are a consequence of a post-translational process by the action of an asparaginyl endoprotease (Egidi et al. 2014). This enzyme cleaves the first three residues of the N-terminal domain (MEN- or IEN-), resulting in a mature protein with Ser as the first amino acid. Egidi et al. (2014) suggested that the presence or absence of the three extra residues present in LMW-m mature subunits (MET-), which are unaffected by the asparaginyl endoprotease, likely does not imply a structural difference and would not affect gluten properties. The mutations found in N-terminal for two LMW-s genes of Group-III (Asn3 $\rightarrow$ Ser) could lead to differential post-translational processing, without the elimination of the three first residues of this domain, releasing a novel LMW-s with an N-terminal domain not previously described.

In conclusion, a large number of novel LMWGs alleles were detected in the *Sitopsis* species evaluated, with 18 novel active variants all from LMW-s subunits not associated with genes from common wheat, as well as the LMW-m of the *pGluU* group, which could be useful in deriving products suitable for celiac patients. These results show that the *Sitopsis* species are an excellent source of novel variability that could be used in breeding programmes for quality improvement in common wheat and for products with advantages in relation to celiac disease.

Acknowledgements

This research was supported by Grant AGL2014-52445-R from the Spanish Ministry of Economy and Competitiveness, co-financed by the European Regional Development Fund (FEDER) from the European Union; and Grant P11-AGR-7920 from the Regional Government of Andalusia (Southern Spain). The first author is grateful to the Spanish Ministry of Economy and Competitiveness (FPI programme) and European Social Fund for a predoctoral fellowship.

CAPÍTULO IV

CARACTERIZACIÓN MOLECULAR DE NUEVAS SUBUNIDADES LMW-i DE GLUTENINA EN *Triticum urartu* Thum. ex Gandil.

Publicado como:

S. Cuesta, C. Guzmán, J.B. Alvarez (2015) Molecular characterization of novel LMW-i glutenin subunit genes from *Triticum urartu* Thum. ex Gandil. *Theoretical and Applied Genetics* **128**: 2155-2165.

Resumen

Las subunidades de bajo peso molecular de glutenina son importantes en la determinación de las propiedades viscoelásticas de la masa de harina de trigo. *Triticum urartu* Thum. ex Gandil., el cual está relacionado con el genoma A de los trigos poliploides, ha mostrado ser una buena fuente de variación para estas subunidades. El presente estudio evaluó la variabilidad de los genes de LMW-i en esta especie. Un alto polimorfismo fue encontrado en las secuencias analizadas y dio lugar a la detección de 11 nuevos alelos, clasificados en dos grupos (Grupo I y –II) que mostraron SNPs y InDels únicos. Ambos grupos estuvieron asociados con genes de *Glu-A3-1* de trigo común. En general, las proteínas deducidas de los genes del Grupo II poseían una alta proporción de glutamina y prolina, lo cual ha sido sugerido previamente estar relacionado con buena calidad. Además, hubo algunos cambios respecto a trigo común. Esta nueva variación podría afectar a la calidad de la masa. Epítopos adicionales para la enfermedad celíaca fueron también detectados, sugiriendo que estas subunidades podrían ser altamente reactivas. Los resultados mostraron que *T. urartu* podría ser una importante fuente de variabilidad genética para los genes de LMW-i que podrían aumentar el acervo genético del trigo moderno.

Palabras clave: enfermedad celíaca, recursos genéticos, calidad del gluten, genes de LMW-i, caracterización molecular, *Triticum urartu*.

Abstract

Low molecular weight glutenin subunits (LMWGs) are important in determining the viscoelastic properties of wheat dough. *Triticum urartu* Thum. ex Gandil., which is related to the A genome of polyploid wheat, has been shown as a good source of variation for these subunits. The present study evaluated the variability of LMW-i genes in this species. High polymorphism was found in the sequences analysed and resulted in the detection of 11 novel alleles, classified into two sets (Group-I and -II) showing unique SNPs and InDels. Both groups were associated with *Glu-A3-1* genes from common wheat. In general, deduced proteins from Group-II genes possessed a higher proportion of glutamine and proline, which has been previously suggested to be related with good quality. Moreover, there were other changes compared to common wheat. This novel variation could affect dough quality. Additional epitopes for celiac disease were also detected, suggesting that these subunits could be highly reactive. The results showed that *T. urartu* could be an important source of genetic variability for LMW-i genes that could enlarge the genetic pool of modern wheat.

Key words: celiac disease, genetic resources, gluten quality, LMW-i genes, molecular characterization, *Triticum urartu*.

Introduction

The bread-making quality of wheat flour has been widely associated with the endosperm storage proteins (see Wrigley et al. 2006 for reviews). Of these, two main groups, gliadins and glutenins, are the main components of gluten, the protein network that gives the dough its viscoelastic properties (elasticity and extensibility). Both groups differ in their molecular characteristics (Payne 1987). Glutenins are classified in high- (HMWGs) and low-molecular-weight (B-LMWGs and C-LMWGs) subunits. The HMWGs are the polypeptides mainly responsible for bread-making quality in common wheat, whereas this role is largely assumed by the B-LMWGs in durum wheat for pasta-making quality (Wrigley et al. 2006). The HMWGs have been most extensively studied because of their relatively small number and the ease of characterization by SDS-PAGE, whereas the characterization of LMWGs has been more difficult to elucidate due to the complexity of this multigene family (D'Ovidio and Masci 2004). The B-LMWGs are encoded by *Glu-3* loci located on the short arms of chromosomes 1A, 1B and 1D (Singh and Shepherd 1988; Pogna et al. 1990), which are closely linked to the *Gli-1* loci for γ- and ω-gliadin genes. The B-LMWGs have been divided into three types: LMW-i (isoleucine), LMW-m (methionine) and LMW-s (serine) depending on the first amino acid of the mature protein (D'Ovidio and Masci 2004). The LMW-i are only coded by the *Glu-A3* locus whereas the LMW-m are synthesised by the three loci and the LMW-s are coded by *Glu-B3* and *Glu-D3*.

The standard LMWG structure consists of four main domains: signal peptide, N-terminal, repetitive and C-terminal. The latter domain is further subdivided into three regions (Cassidy et al. 1998): the cysteine-rich (I), glutamine-rich (II) and highly conserved (III) domains. However, the LMW-i subunits possess a unique structure compared to LMW-m and LMW-s, because the N-terminal domain is missing. The LMWGs usually contain eight cysteine residues: seven in the C-terminal domain, and one in a variable position, in the N-terminal or repetitive domains for the LMW-m and LMW-s or in the C-terminal for the LMW-i (Ikeda et al. 2002). The first (or third for LMW-i) and seventh cysteine residues are involved in inter-molecular disulfide bonds, while the remainder form three intra-molecular disulfide bonds (D'Ovidio and Masci 2004). Consequently, in LMW-i the repetitive domain is precluded in the formation of the inter-

molecular disulfide bonds in the gluten polymer. This different structure could lead to functional differences with respect to LMW-m and LMW-s and have a different impact on the viscoelastic properties of dough (Cloutier et al. 2001; Ikeda et al. 2002). Several studies have also associated LMWGs with celiac disease due to the presence of reactive epitopes as for those exhibited in gliadins (see Rasheed et al. 2014 for reviews).

A recent study showed that there were at least 15 LMWGs genes in individual accessions of common wheat (Zhang et al. 2013). The *Glu-A3* locus showed the highest allelic diversity and *Glu-B3* showed moderate diversity, whereas *Glu-D3* was very low. For the *Glu-A3* locus, Wang et al. (2010) identified up to three LMWG genes (*Glu-A3-1*–*Glu-A3-3*), the former encoding LMW-m and the latter two encoding LMW-i. However, Zhang et al. (2013) identified 4–6 genes (two for LMW-m and the rest for LMW-i) suggesting that these last were organized in different haplotypes.

Additional to the evaluation and characterization of the LMWG genes detected in modern wheat, the search for variability to extend the genetic pool is very important for wheat improvement (Jauhar 1993). In this respect, the putative diploid ancestors of the wheat genome could be good sources of useful genes (Srivastava and Damania 1989). As previously mentioned, the LMW-i subunits have been associated with the A genome, whose ancestor has been identified as *Triticum urartu* Thum. ex Gandil. (Dvorak et al. 1993), a wild wheat species (2n = 2× = 14; A^uA^u) of the Fertile Crescent region (Johnson 1975; Miller 1987). Studies by protein separation in SDS-PAGE have shown a wide polymorphism for endosperm storage protein in this species (Rodríguez-Quijano et al. 1997; Lee et al. 1999b; Caballero et al. 2008; Martín et al. 2008). In our previous study (Caballero et al. 2008), one broad collection (169 accessions) of *T. urartu* was evaluated for variability of HMWGs and B-LMWGs using SDS-PAGE analysis, and 17 HMWGs and 24 B-LMWGs alleles were detected. Of 17 HMWGs alleles, 12 were molecularly characterized by Alvarez et al. (2013), showing differences between these alleles and those present in common wheat. Of the B-LMWGs alleles, 20 in the collection were considered rare or very rare (frequency ≤ 5%).

The aim of the current study was to molecularly characterize the LMW-i glutenin genes present in eight *Glu-A^u3* allelic variants identified in *T. urartu* (*Glu-A3ad*, *Glu-A3af*, *Glu-A3ag*, *Glu-A3ak*, *Glu-A3ao*, *Glu-A3aq*, *Glu-A3au* and *Glu-A3aw*) and to analyse their relationship with those present in common wheat.

Materials and methods

Plant materials

Eight accessions of *T. urartu* previously analysed by SDS-PAGE (Caballero et al. 2008) were used in this study (Table 1). These accessions were obtained from the National Small Grains Collection (Aberdeen, Idaho, USA) and the Institute for Plant Genetics and Crop Plant Research (Gatersleben, Germany).

Protein extraction and electrophoretic analysis and mass spectrometry

Proteins were extracted from single crushed seeds according to the protocol described by Alvarez et al. (2001). Reduced and alkylated glutenin subunits were fractionated by electrophoresis in vertical SDS-PAGE slabs in a discontinuous Tris-HCl-SDS buffer system (pH: 6.8/8.8) at a polyacrylamide concentration of 10% (w/v, C: 1.28). The Tris-HCl/glycine buffer system of Laemmli (1970) was used. Electrophoresis was carried at 30 mA/gel and 18°C for 45 min after the tracking dye migrated off the gel. Gels were stained overnight with 12% (w/v) trichloroacetic acid solution containing 5% (v/v) ethanol and 0.05% (w/v) Coomassie Brilliant Blue R-250. De-staining was carried out with tap water.

At the same time, cold acetone was added to the same sample supernatants, and then the LMWGs were allowed to precipitate. The samples were then used for matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF-MS) on an AB Sciex 5800 TOF-TOF apparatus (AB Sciex, Darmstadt, Germany). The matrix used was α-ciano-4-hidroxicinámico (CHCA). The calibration was done with the calibration kit Cal Mix3 (AB Sciex, Darmstadt, Germany). According to the molecular weights of LMWGs obtained from gel electrophoresis and mass spectrometry, the corresponding protein subunit encoded by the studied genes was identified

DNA extraction and PCR amplification

Genomic DNA was isolated from young leaves of a single plant per accession according to the method of CTAB (Stacey and Isaac 1994). In order to amplify the complete coding region of the LMW-i genes primers 5'-ATGAAGACCTTCCTCGTCTTT-3' (Ma et al. 2006) and 5'-TCACACATGACGTTGTGTGAC-3' (Zhang et al. 2011) were used. PCR amplification of genomic DNA was performed in a volume total of 20μl containing 50 ng of genomic DNA, 0.3 μM of each primer, 0.4 mM of dNTPs, 1 mM or 1.5 mM of $MgCl_2$, 1× of

reaction buffer and 1 U of *Taq* DNA polymerase (Promega, Madison, WI, USA). The amplification was carried out with a first step of initial denaturation at 94 °C for 3 min followed by 35 cycles of 30 s of denaturation at 94 °C, a step of annealing of 30 s at 58 °C or 60 °C and then 1.5 min of extension at 72 °C. To finish the process, an extension final step at 72 °C for 10 min was performed. The PCR products (amplicons) were separated by electrophoresis on polyacrylamide gels of 8% (w/v, C: 1.28%), stained with ethidium bromide and visualised under UV light.

DNA sequencing analysis

The PCR products were purified using Sureclean (Bioline) and then ligated to pGEM-T (Promega, Madison, WI, USA) and used to transform *Escherichia coli* JM109 competent cells. Thirty colonies of each cloned PCR product were analysed for the presence of LMWGs genes insert. Inserts were amplified with M13 universal primers (binding region adjacent to the insert in the vector) and the PCR products were separated using electrophoresis on polyacrylamide gels of 8% (w/v, C: 1.28%). At least three different inserts were sequenced using an ABI Prism 310 Genetic Analyzer (Applied Biosystems, Carlsban, CA, USA). The novel sequences are available from Genbank database.

Data analysis

The sequences obtained were analysed and compared using the Geneious Pro ver. 5.0.3 software (Biomatters Ltd.). Phylogenetic tree was constructed with MEGA5 software (Tamura et al. 2011) using the complete coding sequences obtained together with the sequences of the LMW-i genes identified in common wheat by Zhang et al. (2013), also as the LMW-i gene isolated from common wheat cv. Chinese Spring (NCBI: AY453154; Zhang et al. 2004) and cv. Glenlea (NCBI: AY542896; Cloutier et al. 2001). Six sequences isolated from einkorn (*T. monococcum* L. ssp. *monococcum*; 2n = 2× = 14, A^mA^m) were also included in this comparison (NCBI: AY146588-2 and AY146588-3, Wicker et al. 2003; NCBI: DQ307388, DQ307389 and DQ345449, An et al. 2006; and NCBI: DQ234068, Ma et al. 2006). Neighbour-joining cluster with all sequences analysed was generated using the Maximun Composite likelihood method (Tamura et al. 2004) and one bootstrap consensus from 1000 replicates was used (Felsenstein1985).

DNA analyses were conducted by DnaSP ver. 5.0 (Librado and Rozas 2009) and parameters as total number of mutations (η), average number of nucleotide differences (k) and number of polymorphic sites(s) were calculated. Nucleotide diversity was estimated as theta (θ), the number of segregating (polymorphic) sites (Watterson 1975), and pi (π), the average number of nucleotide differences per site between two sequences (Nei 1987). Tests of neutrality were performed using Tajima's D statistic (1989).

Results

Isolation and variation of LMWGs genes

Eight of 24 *Glu-A3* alleles described by Caballero et al. (2008) using SDS-PAGE electrophoresis were analysed by PCR amplification: four were catalogued as rare (frequency ≤ 5%), and the other four were very rare (frequency ≤ 1%). Six of them (three rare and three very rare) were detected in Turkish accessions, while the other two, one rare and other very rare, were of Lebanese and Iraqi origin, respectively (Table 1). These alleles are formed by several protein components (1-3 bands, Fig. 1a). In the current analysis, the specific amplification of the LMWG genes from these accessions revealed several amplicons (Fig. 1b), although some of them could be LMW-m genes, which were not the main subject of this study, and not LMW-i. Thus, 1-3 LMWG sequences were identified for each accession. In total we sequenced 15 sequences (Table 1), but some of them were the same. As the result, there are 11 different sequences with a size range of 894-1062 bp, among which seven were active genes and four were pseudogenes (KJ780779, KJ780780, KJ780783/KJ801156 and KJ780787) because they had 1-3 in-frame stop codons. The presence of these pseudogenes made difficult to establish a univocal relation between the subunits observed in SDS-PAGE (Fig. 1a) and the amplicons detected (Fig. 1b). For the *Glu-A^{u}3af* and *Glu-A^{u}3aq* (Fig. 1, lanes 1 and 3), all the LMW-i sequences evaluated were pseudogenes, while for the rest of the alleles the sequences analysed showed very similar sizes (Table 1), being consequently difficult separated in the gel. For this reason, some sequences have been associated to one unique band in the Fig. 1b. The deduced amino acid sequences of all had an isoleucine as the first residue in the mature protein (codon 61-63) and were classified as LMW-i genes.

Table 1. LMW-i sequences in *T. urartu*.

Accession (Allele[a])	Frequency[b]	NCBI ID:	DNA size (bp)	Mr (kDa)		Group	Repetitive size (aa)	Motifs		Glu +Pro (%)	
				Deduced[c]	MALDI[d]			Number	Size	Repetitive	Total
Iraqui accession											
PI 428253 (*Glu-A3ao*)	vr	KJ780780	1050	N	N	I	-	-	-	-	-
		KJ780787	894	N	N	II	-	-	-	-	-
		KP793238	1059	38.43	38.76	II	162	20	6-14	72.8	53.7
Lebanese accession											
PI 428328 (*Glu-A3ag*)	R	KJ780782	1056	38.23	37.29	I	149	19	6-9	72.5	52.9
		KJ780784	1059	38.43	39.12	II	162	20	6-14	72.8	53.7
Turkish accessions											
PI 428186 (*Glu-A3ak*)	R	KJ780778	1047	37.81	38.33	I	149	19	6-9	73.1	52.7
		KJ780785	1062	38.56	38.33	II	162	20	6-14	72.8	53.7
		KJ780786	1059	38.43	38.33	II	162	20	6-14	72.8	53.7
PI 428188 (*Glu-A3ad*)	R	KJ780779	1050	N	N	I	-	-	-	-	-
		KP793237	1059	38.43	38.26	II	162	20	6-14	72.8	53.7
PI 428191 (*Glu-A3au*)	vr	KJ780781	1044	37.72	38.91	I	149	19	6-9	72.4	52.3
		KJ780788	1056	38.33	38.91	II	155	19	6-16	74.2	54.7
PI 428225 (*Glu-A3aq*)	vr	KJ801156	1050	N	N	I	-	-	-	-	-
PI 428255 (*Glu-A3af*)	R	KJ780783	1050	N	N	I	-	-	-	-	-
TRI 11496 (*Glu-A3aw*)	vr	KJ801157	1059	38.43	38.54	II	162	20	6-14	72.8	53.7

r: rare (≤ 5%), vr: very rare (≤ 1%).

[a] McInstosh et al. 2013.

[b] Classification according frequency in the original collection (Caballero et al. 2008).

[c] N: pseudogene. Data of the mature proteins without signal peptide.

[d] Molecular weight by MALDI-TOF-MS analysis.

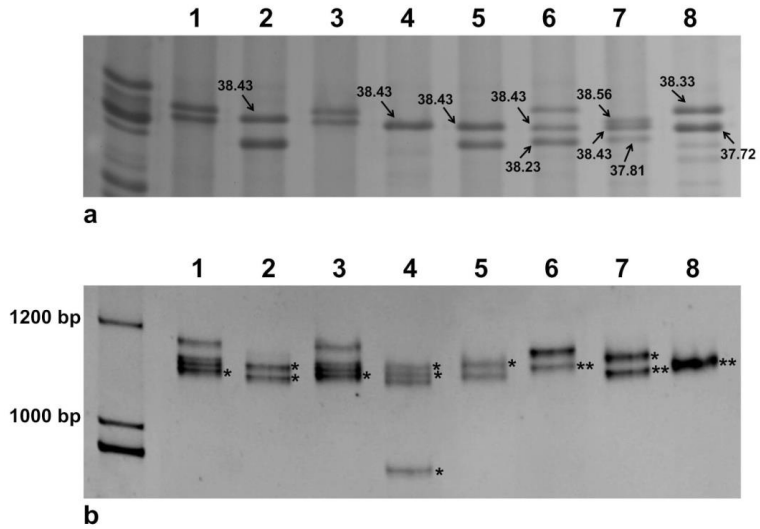


Figure 1. SDS-PAGE separation of LMW-i subunits (**a**) and PCR amplification of LMW-i glutenin genes (**b**) from *T. urartu*. Lane as follows: 1, PI 428255; 2, PI 428188; 3, PI 428225; 4, PI 428253; 5, TRI 11496; 6, PI 428328; 7, PI 428186; and 8, PI 428191. Each asterisk indicates one amplicon sequenced. Numbers near each protein band indicate deduced molecular weights (kDa) based on the amplicon sequenced information.

The 11 genes obtained here were analysed together with other LMW-i genes from einkorn and common wheat to evaluate the relationships between them (Fig. 2). The sequences were grouped into two clusters that corresponded with the *GluA3-1* and *GluA3-3* genes described by Wang et al. (2010). There were different groups within each cluster. The *GluA3-1* gene had up to three sets: the first one associated with *A3-640/A3-649-2* genes described by Zhang et al. (2013); the second with the *A3-573/A3-620/A3-646a/A3-646b/A3-649-1* genes, and the third with *A3-565/A3-568/A3-662* - named in a previous study as *A3-2*, *A3-3* and *A3-4*, respectively (Dong et al. 2010). The other cluster (*GluA3-3*) showed two groups associated with *A3-484* and *A3-502*, respectively. The *T. urartu* sequences were grouped in two sets that we named Group-I and -II inside the *GluA3-1* cluster (Fig. 2). Group-I of *T. urartu* appeared associated with the *A3-640/A3-649-2* sequences, whereas Group-II showed more similarity with the *A3-573/A3-620/A3-646a/A3-646b/A3-649-1* sequences. All LMWi einkorn sequences appeared clearly separated to the *T. urartu* sequences (Fig. 2). Five of them formed a differentiate cluster, while the other one (LMW-M1) was associated with the A3-662 sequence described by Zhang et al. (2013).

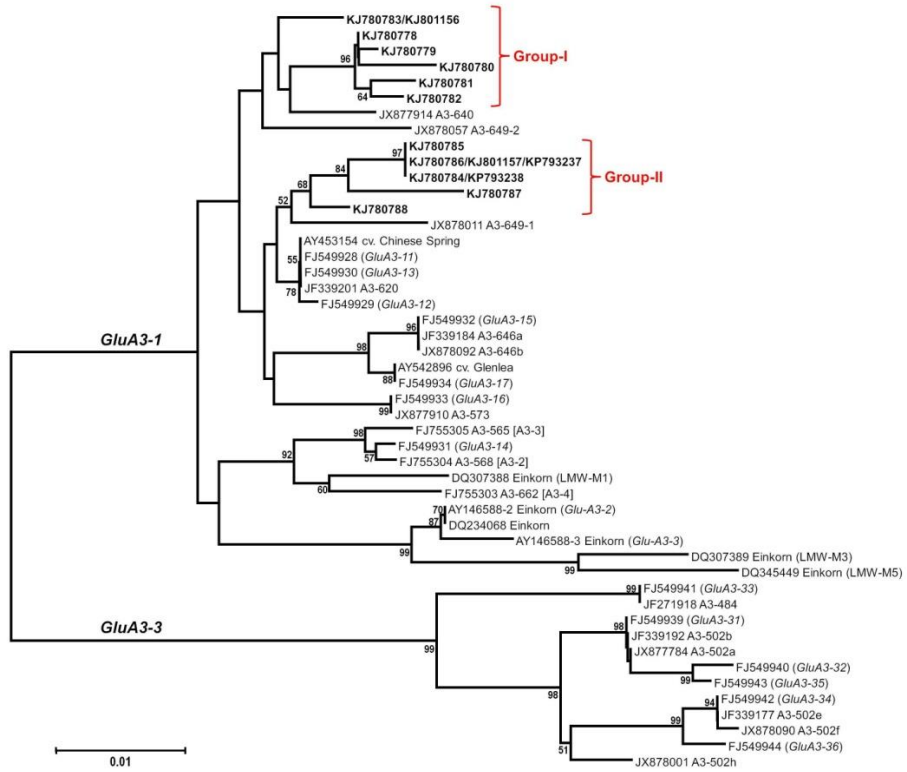


Figure 2. Neighbour-joining tree based on the maximum composite likelihood method of LMW-i gene sequences detected in *T. urartu* accessions (bold), together with the previous sequences described in einkorn and common wheat. The denomination used by these last authors appears between *brackets*. Numbers in nodes indicate bootstrap estimates from 1000 replications.

All Group-I sequences had a deletion of 45 bp inside the repetitive domain along with a 21-bp insertion in C-terminal domain III. For Group-II, there were up to three different types. One of them had a deletion of 6 bp in the repetitive domain, whereas the others had two deletions (3 and 24 bp) inside the same domain. A third type was characterized by a long deletion of 177 bp covering the end of the repetitive domain and the beginning of the C-terminal domain.

Table 2. DNA polymorphism and test statistics for selection of 11 sequences from *T. urartu.*

			C-terminal			
	Signal peptide	Repetitive	I	II	III	Complete
GroupI sequences						
Region	1-60	61-507	508-732	733-891	892-1056	1-1056
Size	60	447	225	159	165	1056
η	1	15	5	3	1	25
k	0.53	5.13	1.67	1.00	0.33	8.67
s	1	14	5	3	1	24
SS	0	4	1	2	1	8
NSS	1	11	4	1	0	17
h	2	5	5	4	2	6
$\theta \times 10^{-3}$	7.3	13.7	9.7	8.9	2.7	10.1
$\pi \times 10^{-3}$	8.9	11.5	7.4	6.8	2.0	8.3
D	0.850	-1.347	-1.337	-1.233	-0.933	-1.313
Group II sequences						
Region	1-60	61-552	553-777	778-939	940-1083	1-1083
Size	60	492	225	162	144	1083
η	1	7	3	6	2	19
k	0.40	3.20	1.40	2.60	0.80	8.40
s	1	7	3	6	2	19
SS	1	3	2	4	0	10
NSS	0	4	1	2	2	9
h	2	3	3	3	3	3
$\theta \times 10^{-3}$	8.0	9.8	7.4	20.0	6.7	10.3
$\pi \times 10^{-3}$	6.7	9.4	7.2	18.1	5.6	9.5
D	-0.816	-0.338	-0.175	-0.668	-0.972	-0.582
			C-terminal			
	Signal peptide	Repetitive	I	II	III	Complete
Overall sequences						
Region	1-60	61-552	553-777	778-939	940-1104	1-1104
Size	60	492	225	162	165	1104
η	2	18	9	10	6	45
k	0.51	5.74	2.62	3.49	2.87	15.24
s	2	17	9	10	6	44
SS	1	6	4	5	3	19
NSS	1	12	5	5	3	26
h	3	8	8	7	5	9
$\theta \times 10^{-3}$	11.4	17.4	15.8	23.7	14.2	17.2
$\pi \times 10^{-3}$	8.5	17.3	13.4	24.2	20.0	17.4
D	-0.778	-0.293	-0.622	-0.096	1.593	-0.039

η: total number of mutations; *k*: average number of nucleotide differences; *s*: number of polymorphic sites; SS: synonymous substitutions; NSS: non-synonymous substitutions; *h*: number of haplotypes; θ: Watterson's estimate; π: nucleotide diversity; and *D:* Tajima's estimate *D*-test.

Comparison of the complete sequences showed the highest level of polymorphism in Group-I sequences, with 25 mutations at 24 polymorphism sites and 17 non-

synonymous changes (Table 2). For Group-II sequences, there were 19 polymorphic sites with nine non-synonymous changes (Table 2). For all sequences (Group-I and -II), the total number of polymorphism sites and mutations were higher than those of individual groups, with 44 and 45, respectively; of those polymorphic sites 26 were non-synonymous changes (Table 2). The repetitive domain showed the largest degree of variation for all sequences and for each individual group, followed by C-terminal II for overall sequences and the Group-II alleles. For the Group-I sequences, the second-most variable domain was the C-terminal I, with five mutations and polymorphism sites. The most conserved domain was the C-terminal III for overall sequences.

Assessment of the nature and function of the seed storage proteins determined that these genes were evolutionarily neutral. In this respect, estimation of the nucleotide diversity of the DNA sequences, obtained here by two statistics - pi (π) and theta (θ) - suggested that this diversity was associated with a drift-mutation balance, consistent with a neutral equilibrium shown by a non-significant Tajima's D-test (Table 2).

Deduced amino acid sequence analysis

The size of deduced mature proteins of Group-I were 327-331 amino acid residues with the deduced molecular weights of 37.72-38.23 kDa in size, while those of Group-II were 331-332 amino acid residues with the deduced molecular weights of 38.33-38.56 kDa (Table 1). The relation between these deduced proteins and the bands detected by SDS-PAGE (Fig. 1a) were established by the further analysis with MALDI-TOF-MS. As shown in Supplementary Fig. 1, the mass spectrometric data detected apparent protein peaks within of each accession with molecular weights consistent with that from deduced protein (Table 1). Only in two cases (lane 7: PI 428186; and lane 8: PI 428191), the MALDI-TOF-MS analysis are not shown a clear discrimination among the different subunits. Both profiles showed one unique peak associated with three or two subunits evaluated, respectively (Supplementary Fig. 1b). The small differences observed between both Mr data could be result of post-translational modification as other authors have reported (Lauriere et al. 1996; An et al. 2006).

The additional analysis of deduced sequences showed that the length of the repetitive domain was 155-162 residues in Group-II and 149 in Group-I sequences. The Group-I sequences had 19 repeat motifs in this domain with a range of 6-9 residues, whereas the Group-II had large motifs with up to 14 residues, with the exception of

KJ780788 with a motif of 16 residues. This last sequence had only 19 motifs, while all other Group-II sequences had 20 (Table 1).

Analysis of these repeat motifs showed the presence in these sequences of peptides (Glt-156: PFSQQQQSPF, and Glt-17: PFSQQQQQ), considered as stimulating epitopes of T-cells in celiac disease by Vader et al. (2002). The Glt-156 peptide was detected twice in all these sequences, with the exception of KJ780788 which was detected once. In addition, the Glt-17 peptide was found once in overall sequences, but KJ780782 did not present this motif (Supplementary Fig. 2).

The glutamine and proline content in mature protein was higher in Group-II than in Group-I, with means of 53.9 and 52.6%, respectively; this rate was conserved when only the repetitive domain was analysed, with 73.1 and 72.7%, respectively. The highest glutamine and proline content in overall sequences was for the KJ780788 sequence, with 53.4% for mature protein and 73.0% for the repetitive domain; and the lowest was for KJ780781, with 52.4 and 72.4%, respectively (Table 1).

In the current study, two LMW-i subunits previously detected in two common wheat cultivars (Chinese Spring and Glenlea) were used to compare amino acid sequences. The AY453154 sequence of cv. Chinese Spring showed a high similarity with the AY542896 sequence of cv. Glenlea. Both sequences showed nine amino acid changes: six in the repetitive domain and one in each C-terminal, together with two InDels in the repetitive domain and one in C-II terminal (Supplementary Fig. 3). Comparison of the deduced mature protein of the *T. urartu* sequences and these two LMW-i subunits is shown in Table 3.

The first two positions indicated appeared as one InDel inside the AY453154 sequence, with four of the other 17 changes similar between the last sequence and the overall *T. urartu* sequences. There was no change in the cysteine backbone (eight residues) characteristic of these LMW-glutenin subunits. However, there were some changes surrounding the cysteine residues: Arg234 $\rightarrow$ Gln and Arg342 $\rightarrow$ Thr in overall subunits, and Ala213 $\rightarrow$ Val in the KJ780781 sequence.

Proline and glutamine residues are important in maintenance of LMWG structures. Five replacements affected proline in the subunits of Group-I. In three positions (7, 24 and 287), the existing proline was changed for another residue (Ala, Ser and Gln, respectively), while the other two replacements led to a proline residue in position 63. In two positions (78 and 301), glutamine was changed for another residue (Table 3). For Group-II, only two changes affected proline residues: Pro287 $\rightarrow$ Gln and

Ser112 → Pro; while there were changes in glutamine residues in four positions: 12, 21, 178 and 299 (Table 3).

Table 3. Amino acid comparison of two LMW-i subunits from common wheat and the Group-I and -II sequences detected in *T. urartu.*

	Repetitive domain								
Position*	7	12	21	24	27	44	63	78	112
AY453154 (cv. Chinese Spring)	-	-	Gln	Pro	Ser	Leu	Pro	Gln	Pro
AY542896 (cv. Glenlea)	Pro	Gln	Gln	Pro	Ser	Ser	Leu	Gln	Ser
Group I									
KJ780778	Ala					Ala	Pro		-
KJ780781	Ala			Ser	Leu	Ala	Pro		-
KJ780782	Ala				Leu	Ala	Pro	His	-
Group II									
KJ780784/KP793238		Glu	Lys				Pro		Pro
KJ780785		Glu	Lys				Pro		Pro
KJ780786/KJ801157/KP793237		Glu	Lys				Pro		Pro
KJ780788		Glu					Pro		Pro

	C-I terminal			C-II terminal				C-III terminal		
Position*	178	213	234	256	287	299	301	324	339	342
AY453154 (cv. Chinese Spring)	Gln	Ala	Gln	Ile	Gln	Gln	Gln	His	Thr	Thr
AY542896 (cv. Glenlea)	Gln	Ala	Arg	Ile	Pro	Gln	Gln	His	Thr	Arg
Group I										
KJ780778			Gln	Val	Gln		Glu		Asn	Thr
KJ780781		Val	Gln	Val	Gln		Glu		Asn	Thr
KJ780782			Gln	Val	Gln		Glu		Asn	Thr
Group II										
KJ780784/KP793238	Arg		Gln		Gln	Leu		Tyr		Thr
KJ780785	Arg		Gln		Gln	Leu		Tyr		Thr
KJ780786/KJ801157/KP793237	Arg		Gln		Gln	Leu		Tyr		Thr
KJ780788			Gln		Gln	Leu				Thr

*The positions are determined by the mature protein of the AY542896 sequence (LMWG-50) isolated by Cloutier et al. (2001).

In Addition to these amino acid changes, some important InDels mainly in the repetitive and C-III domains were detected in these sequences of *T. urartu* (Supplementary Fig. 3). All sequences showed one extended deletion in the repetitive domain, being 38 residues in Group-I sequences and 23 in Group-II. Furthermore, the KJ780788 sequence presented one additional deletion (PPFSQQQQ) between residues 171 and 178 of the reference sequence (AY542896). For Group-I, two insertions of one glutamine residue were detected: one in the repetitive domain (residues 110-111 of AY542896) and other in the C-II domain (residues 276-277). This first insertion was also detected in all Group-II sequences, while the latter was not found in the KJ780788 sequence, which had a unique deletion (HGTFLQP) in the C-III domain (Supplementary Fig. 3).

Discussion

The low variation of some traits of interest in wheat breeding has suggested the possibility of searching for variation in wheat relatives. Among them, the species identified as putative donors of the three genomes present in common wheat could be the main candidates, because they are included in the primary gene pool and their crossing with wheat does not require special techniques.

Numerous studies have suggested that *T. urartu* is the species donor of the A genome (Dvorak et al. 1993), an event that likely occurred ~0.5 million years ago (Huang et al. 2002; Dvorak and Akhunov 2005). Among the traits related to wheat quality that could be transferred from this wild diploid wheat to modern wheat are seed storage proteins (glutenins and gliadins). A particular characteristic of the A genome is the presence of a LMWG type (LMW-i) not detected in the other wheat genomes. These glutenin subunits are exclusively synthesised by genes of the *Glu-A3* locus, consequently *T. urartu* is the natural source of these subunits. In this study, seven genes and four pseudogenes of this LMWG type were detected and sequenced in eight *T. urartu* accessions.

Recent studies suggested that the LMWG genes in the A genome are synthesised by two genes for the LMW-m glutenin subunits and 2-4 genes for the LMW-i glutenin subunits (Zhang et al. 2013). According to previous classifications (Wang et al. 2010), the LMW-m subunits are encoded by the *GluA3-2* gene, and LMW-i by the *GluA3-1* and *GluA3-3*. The reconciliation of this classification with the data of Zhang et al. (2013) and data of the present study suggests that the *GluA3-2* gene of Wang et al. (2010) corresponds with the m_{AD} cluster defined by Zhang et al. (2013), which was formed by two genes.

The *GluA3-3* gene could be also constituted by two genes: one most frequent formed by all variants of the *A3-502* gene (Zhang et al. 2013), together with all alleles of this gene described by Wang et al. (2010), without the *GluA3-33* allele that appeared associated with the other gene (*A3-484*), which is less frequent (Zhang et al. 2013). This latter sub-group of the *GluA3-3* gene formed the i_A-5 haplotype according to Zhang et al. (2013), together with one sub-group of the *GluA3-1* gene constituted by the *A3-565/A3-568/A3-662* genes, previously described as *A3-2*, *A3-3* and *A3-4* by Dong et al. (2010). In this last group could also be included the LMW-i genes found in *T. monococcum* L. ssp. *monococcum* by Wicker et al. (2003) and Ma et al. (2006). This suggests that this group

could have a different origin to that indicated by these authors, and this should be studied in the future.

Similar to the other groups found, the LMW-i haplotype (i_A) could be generally formed by up to three genes: one or two for *Glu-A3-1* and one for *GluA3-3*, being the two sequence sets of *T. urartu* obtained in this study associated with both genes of *GluA3-1*. Nucleotide diversity was high, which could have two main causes: the neutral nature of these genes to evolution (Gepts 1990), and the role of these proteins in plant biology where they are mainly a source of amino acids. These high values are frequent in wild species because lack of selection pressure has generated no advantage of one variant over others. However, in cultivated species, this effect is counteracted by unconscious selection on the part of farmers of the allelic variants due to their function in quality properties of desirable food products. This narrows genetic variability by genetic drift processes (Gepts 1990).

Six of the eight accessions evaluated in this study showed several active LMW-i genes and the relation between their deduced proteins and the bands detected by SDS-PAGE could be established. However, for the other two accessions were not found actives LMW-i genes among the PCR products obtained in current study. Some of these PCR products were identified as LMW-m genes (results not shown), which open the possibility of the expressed proteins of these accessions could be LMW-m, although no LMW-m protein encoded has been identified so far.

Both *T. urartu* groups showed clear differences inside their sequences, mainly in the presence of exclusive SNPs and InDels. These InDels appeared mainly in the repetitive domain, whose structure of tandem motifs could generate duplication or deletion of one or various motifs by slippage during replication (Cassidy and Dvorak 1991). These duplications or deletions may cause repetitive domains to be generally large, which could be associated with dough quality as suggested by Masci et al. (1998, 2000). A large repetitive domain provides a greater number of glutamines available for inter-molecular interactions through hydrogen bonds that could strengthen the gluten network, thus increasing dough elasticity. Furthermore, SNP variations could also be a source of functional changes, particularly those substitutions affecting the structure of LMWGs (Tanaka et al. 2005; An et al. 2006). Although the repetitive domain of all these sequences showed deletions that make these sequences are smaller than those in common wheat, in general, all deduced proteins from Group-II genes possess a longer repetitive domain than for Group-I. Furthermore, several SNPs detected in these *T. urartu* genes,

mainly the proline and glutamine substitutions, could affect gluten strength because this alters protein structure and so affect its elasticity.

Another important feature of LMWGs is the presence of eight cysteine residues, which are important in forming the inter- and intra-molecular disulfide bonds that determine the structure of the gluten polymer (D'Ovidio and Masci 2004). These cysteines were conserved in the 11 LMW-i genes detected in *T. urartu*. However, positions around the cysteine residues are also important in the formation of disulfide bonds and appear to be highly conserved (Masci et al. 1998). Substitutions at this level were found in the present study. Two changes could affect three cysteine residues in all alleles, while the KJ780781 allele could affect five of eight cysteine residues. This replacement could affect protein structure and consequently that of the gluten network.

Although celiac disease has been mainly associated with the gliadins, some authors have found reactive epitopes in the HMWG and LMWG- subunits (van de Wal et al. 1998; Vader et al. 2002). Two epitopes described by Vader et al. (2002) were found in the LMW-i genes detected in *T. urartu* accessions of the present study. Curiously, due to the InDel present in the repetitive domain, these proteins showed one additional motif not present in the LMW-i subunits of common wheat used as a reference, what means that probably could be more reactive. This suggests that although the quality genetic pool of common wheat could be increased by introgression of these proteins, the new materials have no advantage in possible use by celiac patients. On the contrary, some LMW-m subunits detected in wheat do not present these epitopes (results not shown). Although further studies should be carried out, this suggests that the LMW-i subunits could be related to celiac disease of wheat flour and the LMW-m subunits are not.

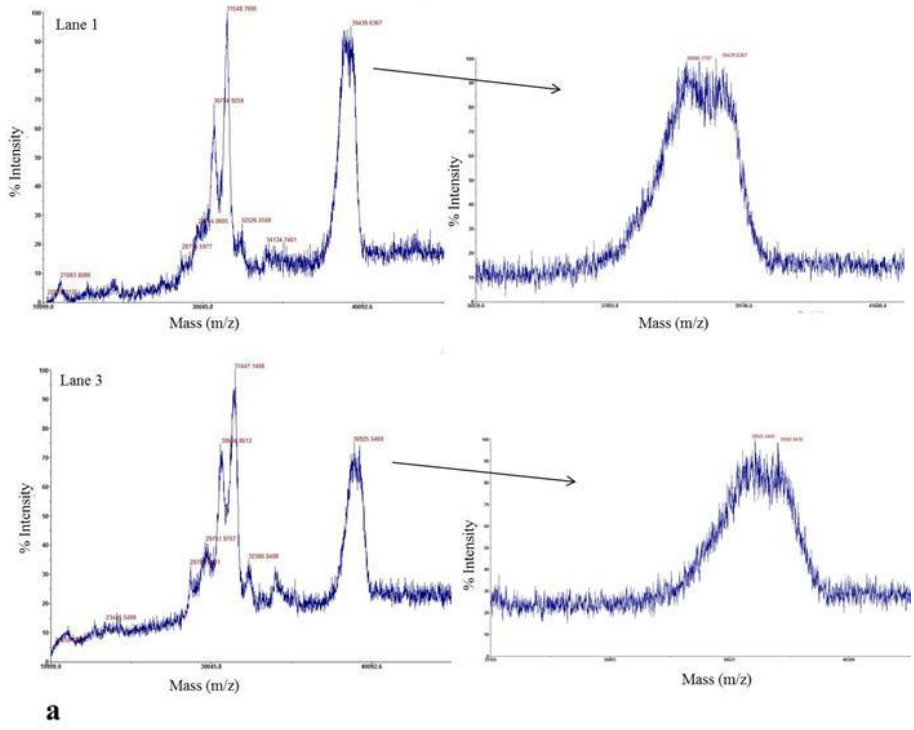
In conclusion, *T. urartu* was shown to be an important source of novel variation for LMW-i genes. All 11 novel genes detected were associated with *Glu-A3-1* genes and showed differences from those in common wheat, which could lead to functional changes. Therefore, these alleles might be useful in wheat breeding for quality improvement. Further studies are required to evaluate the effect of the novel alleles on the quality of modern wheat and their relationship to celiac disease.

Acknowledgements

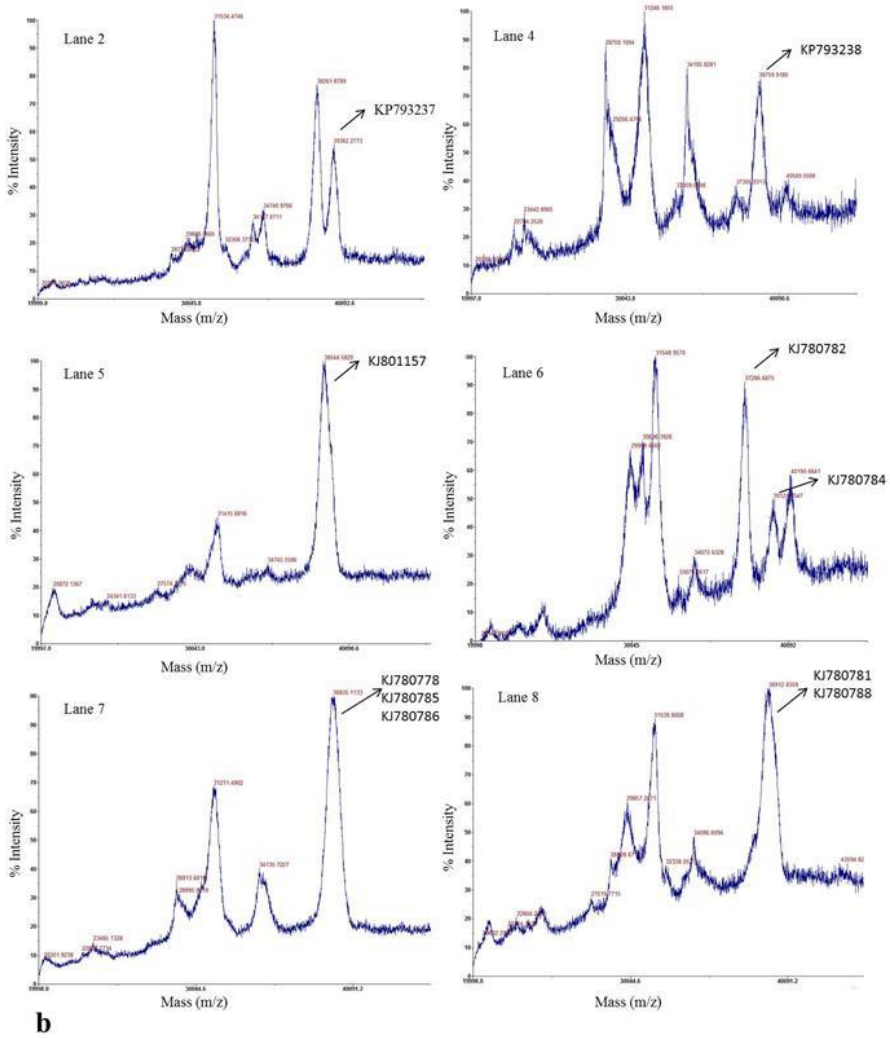
This research was supported by Grant AGL2010-19643-C02-01 from the Spanish Ministry of Economy and Competitiveness, co-financed with European Regional Development Fund (FEDER) from the European Union. The first author is grateful to the

Spanish Ministry of Economy and Competitiveness (FPI programme) and European Social Fund for a predoctoral fellowship. We thank to the National Small Grain Collection (Aberdeen, ID, USA) and the Institute for Plant Genetics and Crop Plant Research (Gatersleben, Germany) for supplying the analysed material. Also, we thank to Dr. E. Chicano (Proteomic Unit, IMIBIC, Córdoba, Spain) for the MALDI-TOF-MS analysis.

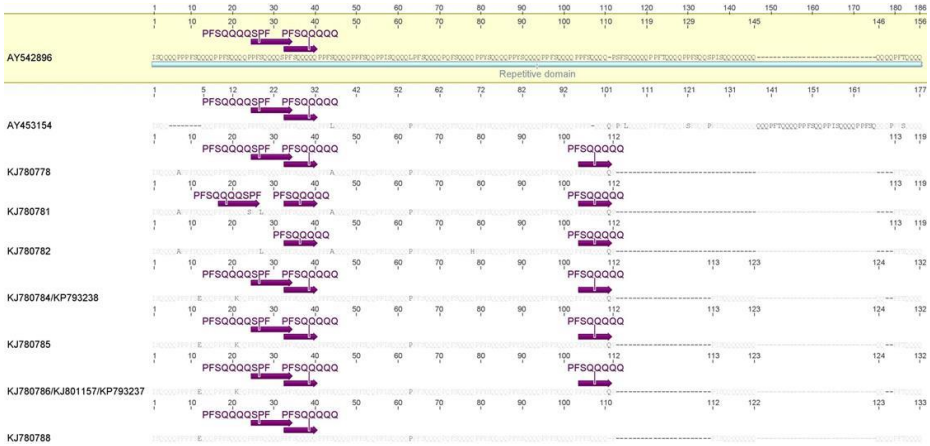
Supplementary material



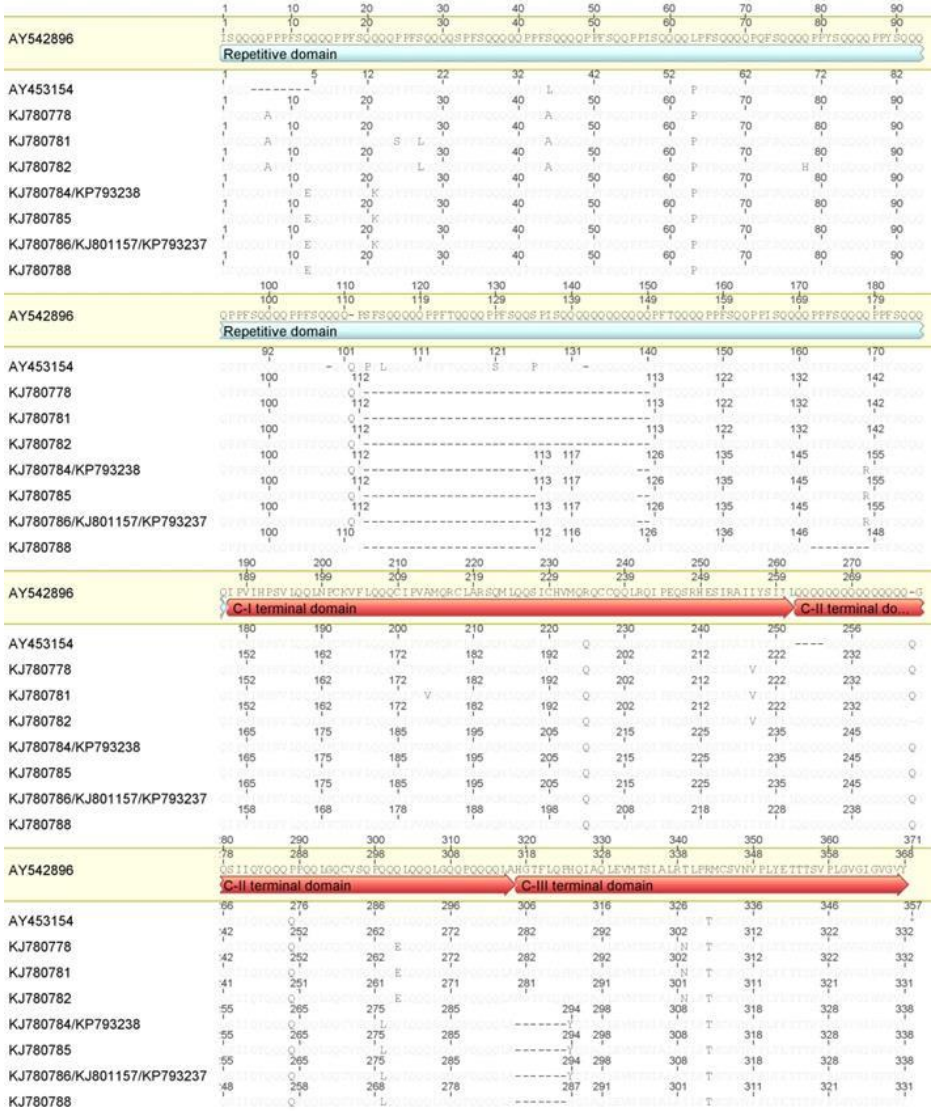
Supplementary Figure 1. The molecular weight (Mr) determination of LMW-i glutenins from *T. urartu* by MALDI-TOF-MS. **a**, profile of the lane 1 (PI 428255) and lane 3 (PI 428225); **b**, profiles of the rest of lanes (2, PI 428188; 4, PI 428253; 5, TRI 11496; 6, PI 428328; 7, PI 428186; and 8, PI 428191). The subunits evaluated appear indicated with their NCBI ID. The lanes corresponded with the SDS-PAGE gel showed in Fig 1a.



Supplementary Figure 1. (Continuation) The molecular weight (Mr) determination of LMW-i glutenins from *T. urartu* by MALDI-TOF-MS. **a**, profile of the lane 1 (PI 428255) and lane 3 (PI 428225); **b**, profiles of the rest of lanes (2, PI 428188; 4, PI 428253; 5, TRI 11496; 6, PI 428328; 7, PI 428186; and 8, PI 428191). The subunits evaluated appear indicated with their NCBI ID. The lanes corresponded with the SDS-PAGE gel showed in Fig 1a.



Supplementary Figure 2. Identification of the Glt-156 (PFSQQQQSPF) and Glt-17 (PFSQQQQQ) epitopes in the LMWi deduced sequences.



Supplementary Figure 3. Alignments of the aminoacid sequences of *T. urartu* with respect to the AY453154 (cv. Chinese Spring) and the AY542896 (cv. Glenlea).

CAPÍTULO V

IDENTIFICACIÓN Y CARACTERIZACIÓN MOLECULAR DE NUEVOS GENES DE SUBUNIDADES LMW-m Y LMW-s DE GLUTENINA Y UN GEN QUIMERA –m/-i DE GLUTENINA EN TRES ESPECIES DIPLOIDES DE *Triticum*

Enviado como:

S. Cuesta, J.B. Alvarez, C. Guzmán (2015) Identification and molecular characterization of novel LMW-m and -s glutenin genes, and a chimeric -m/-i glutenin gene in three diploid *Triticum* species. *Molecular Breeding* (under review).

Resumen

Las subunidades de bajo peso molecular de glutenina (LMWGs) son parte del gluten, el cual da a la masa de harina de trigo sus propiedades viscoelácticas. Las especies diploides con genoma A relacionadas con el trigo común muestran gran variabilidad en estas subunidades. El presente estudio caracterizó la variabilidad de los genes de LMW-m y –s, siendo identificados quince genes nuevos de LMWGs en las tres especies diploides evaluadas. Diez fueron pseudogenes, lo cual es común en las prolaminas de los cereales. Los genes activos correspondieron a los genes de LMW-m y se detectaron algunos polimorfismos de un solo nucleótodo y eventos de inserción/deleccion, lo cual podría alterar la estructura de la proteína y afectar a la calidad de la masa. Dos variantes de los genes de LMW-s fueron detectados en escaña cultivada y *Triticum urartu*. Los genes de LMW-m y -s estuvieron relacionados a los genes *TuA3-391* y *TuA3-400*, y al gen *TuA-460*, respectivamente, de *T. urartu*. El cribado de las secuencias caracterizadas para los epítopos reactivos de la enfermedad celíaca reveló que LMW-m podría ser menos tóxica que otras subunidades para pacientes celíacos. Un nuevo gen quimera con características de genes de LMW-m y de LMW-i fue detectado en *T. urartu*, el cual produce una nueva proteína madura que podría tener un efecto diferente en la calidad de la masa. También fueron proporcionados nuevos conocimientos en la evolución de los genes de LMWG. Estas especies son una potencial fuente de nuevas variantes de genes de LMWGs.

Palabras clave: enfermedad celíaca, genes quimera, escaña, calidad del gluten, genes de LMWG, *T. urartu*.

Abstract

Low molecular weight glutenin subunits (LMWGs) are part of the gluten network that gives dough its viscoelastic properties. A-genome-containing diploid species related to common wheat show great variability in these subunits. The current study characterized the variability of LMW-m and -s genes, being identifying fifteen novel LMWGs genes in the three species evaluated. Ten were pseudogenes, which are common in cereal prolamins. The active genes corresponded to the LMW-m genes and some single nucleotide polymorphisms and insertion/deletion events were detected, which could alter protein structure and affect dough quality. Two variants of the LMW-s genes were detected in cultivated einkorn and *Triticum urartu*. The LMW-m and -s genes were related to *TuA3-391* and *TuA3-400* genes, and to the *Tu-460* gene, respectively, of *T. urartu.* Screening the sequences characterized for reactive epitopes of celiac disease revealed that LMW-m could be less toxic than other subunits for celiac patients. One novel chimeric gene with features from LMW-m and LMW-i genes was detected in *T. urartu*, and it produces a novel mature protein that may have a distinguishing effect on dough quality. Novel insights into the evolution of LMWGs genes are also reported. These species are a potential source of novel LMWGs variants.

Keywords: celiac disease, chimeric genes, einkorn, gluten quality, LMWGs genes, *T. urartu*.

Introduction

Dough's viscoelastic properties, which permit the use of wheat flour to produce diverse foods, are associated with the presence of gluten, a protein network formed by the interactions of two main components: gliadins and glutenins (see Wrigley et al. 2006 for review). Glutenins are further classified into two types of subunits: high-molecular-weight (HMWGs) and low-molecular-weight (LMWGs) based on their mobility in sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The former, encoded by the *Glu-1* locus, have been extensively studied because of their relatively simple pattern. Correlations between allelic variants of *Glu-1* and bread-making quality in common wheat have been established (see Rasheed et al. 2014 for review).

On the contrary, although in durum wheat the LMWGs have been described as the main responsible for the pasta-making quality (gluten strength), their role in bread-making quality has not been clearly established (Wrigley et al. 2006). The LMWGs are formed by a complex mixture containing a great number of subunits encoded by the *Glu-3* loci located on the short arms of chromosomes 1A, 1B and 1D (Singh and Shepherd 1988; Pogna et al. 1990). The compositions of common wheat LMWGs genes were characterized and several genes were identified for each *Glu-3* locus: four for *Glu-A3*, eight for *Glu-B3* and eight for *Glu-D3* (Zhao et al. 2006, 2007; Wang et al. 2009, 2010; Dong et al. 2010 Zhang et al. 2013;). These subunits are classified into three main types according to the first amino acid residue of its mature protein: LMW-i (isoleucine), LMW-m (methionine) and LMW-s (serine) (D'Ovidio and Masci 2004). Furthermore, although LMWGs can be very different in size, they all show a standard structure containing four domains: a signal peptide, N-terminal, repetitive and C-terminal; and the 6+2 cysteine residues can form three intra- and two inter-molecular disulfide bonds, respectively (D'Ovidio and Masci 2004). Variations of this structure can result in functional differences. In previous studies LMWGs allelic forms with unique features were associated with good quality properties (Masci et al 1998; Cloutier et al. 2001; Xu et al. 2006). For this reason, the characterization of novel allelic variants is important in locating genes that can be used to generate new lines with novel quality properties that could be adopted for new uses.

Additionally to the variation found in modern cultivars, the search of new variation in genetic resources has great importance in enlarging the wheat gene pool (Jauhar 1993). Thus, the characterization and comparison of old and new allelic variants is necessary for their effective use. The candidate as variation sources includes both the old wheat varieties and the wheat relatives, mainly those species that donated some of the wheat genomes (Rasheed et al. 2014). Among these species, the main diploid species related to modern wheat is *Triticum urartu* Thum. ex Gandil. (2n = 2× = 14; A^uA^u), which has been identified as the A genome donor in polyploid wheat (Dvorak et al. 1993). Other diploid species containing the A genome (2n = 2× = 14, A^mA^m) are cultivated einkorn (*T. monococcum* L. ssp. *monococcum*) and wild einkorn (*T. monococcum* ssp. *aegilopoides* Link em. Tell.), with the former being domesticated from the latter (Zohary and Hops 1988). The analysis of endosperm storage proteins by SDS-PAGE of these species revealed wide polymorphism (Rodríguez-Quijano et al. 1997; Lee et al. 1999b; Alvarez et al. 2006; Caballero et al. 2008; Martín et al. 2008), which suggests these species could be sources of new variants of the LMWGs genes.

The studies carried out by Wang et al. (2010) and Zhang et al. (2013) suggest that the *Glu-A3* locus is formed by two gene groups: the m_{AD} group with two genes (*Glu-A3-2* and *A3-391*) that encode LMW-m subunits and the i_A group with two other genes (*Glu-A3-1* and *Glu-A3-3*) that encode LMW-i subunits. Based on these studies, the structure of this locus in diploid species from the *Triticum* genus should be similar. In a recent study, eight *Glu-A3* alleles of *T. urartu* characterized by Caballero et al. (2008) were analysed for LMW-i genes displaying a great allelic diversity (Cuesta et al. 2015). In a recent study carried out using *T. urartu* (Luo et al. 2015), the presence of the LMW-s genes was detected (*TuA3-460* gene), which had been not previously associated with the A genome. These authors also described four LMW-m (*TuA3-385*, *TuA3-391*, *TuA3-397* and *TuA3-400*) and three LMW-i (*TuA3-502*, *TuA3-538* and *TuA3-576*) genes. In other diploid species such studies have been scarce, and only a few genes have been characterized so far (Wicker et al. 2003; An et al. 2006; Ma et al. 2006).

The main goal of this study was the identification and molecular characterization of LMWGs genes in three diploid species of the *Triticum* genus, together with the analyses of their relationships with genes from polyploid wheat.

Materials and methods

Plant vegetal

Sixteen accessions from three diploid *Triticum* species (six of wild einkorn, three of cultivated einkorn and seven of *T. urartu*) were used in this study (Table 1). All accessions were kindly provided by the National Small Grains Collection (Aberdeen, Idaho, USA) and the Institute for Plant Genetics and Crop Plant Research (Gatersleben, Germany).

Grain protein extraction, SDS-PAGE and MALDI-TOF-MS

Proteins were extracted from single crushed seeds according to the protocol described by Alvarez et al. (2001). Reduced and alkylated glutenin subunits were fractionated by electrophoresis in vertical SDS-PAGE slabs in a discontinuous Tris-HCl-SDS buffer system (pH: 6.8/8.8) at a polyacrylamide concentration of 10% (w/v, C: 1.28). The Tris-HCl/glycine buffer system of Laemmli (1970) was used. Electrophoresis was carried at 30 mA/gel and 18°C for 45 min after the tracking dye migrated off the gel. Gels were stained overnight with 12% (w/v) trichloroacetic acid solution containing 5% (v/v) ethanol and 0.05% (w/v) Coomassie Brilliant Blue R-250. De-staining was carried out with tap water.

At the same time, cold acetone was added to the same sample supernatants, and then the LMWGs were allowed to precipitate. The samples were then used for matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF-MS) on an AB Sciex 5800 TOF-TOF apparatus (AB Sciex, Darmstadt, Germany). The matrix used was α-ciano-4-hidroxicinámico (CHCA). The calibration was done with the calibration kit Cal Mix3 (AB Sciex, Darmstadt, Germany). According to the molecular weights of LMWGs obtained from gel electrophoresis and mass spectrometry, the corresponding protein subunit encoded by the studied genes was identified

DNA extraction and PCR amplification

Genomic DNA was isolated from young leaves of a single plant per accession using the cetyl trimethyl ammonium bromide (CTAB) method according to Stacey and Isaac (1994). In order to amplify the complete coding region of the LMWGs genes, the primers 5'-ATGAAGACCTTCCTCGTCTTT-3' (Ma et al. 2006) and 5'-TCACACATGACGTTGTGTGAC-3' (Zhang et al. 2011) were used to amplify the region spanning the beginning of the coding region and +104bp inside the 3'-UTR. PCR

amplification of genomic DNA was performed in a volume total of 20μl containing 50 ng of genomic DNA, 0.3 μM of each primer, 0.4 mM of dNTPs, 1.5 mM for *T. urartu* accessions and 2 mM for wild and cultivated einkorn accessions of $MgCl_2$, 1× of reaction buffer and 1 U of *Taq* DNA polymerase (Promega). The amplification was carried out with a first step of initial denaturation at 94 °C for 3 min followed by 35 cycles of 30 s of denaturation at 94 °C, a step of annealing of 30 s at 56 °C for *T. urartu* accessions and 54 °C for wild and cultivated einkorn accessions, then 1.5 min of extension at 72 °C. To finish, an extension final step at 72 °C for 10 min was performed. The PCR products (amplicons) were separated by electrophoresis on 1.2% agarose gels, stained with ethidium bromide and visualised under UV light.

The PCR products were ligated into pGEM-T easy vector (Promega) and then transformed into *Escherichia coli* JM109 competent cells. At least three positive clones of each PCR product were sequenced. The novel sequences are available from Genbank database.

Data analysis

The sequences obtained were analyzed and compared using the Geneious Pro ver. 5.0.4 software (Biomatters Ltd.). Phylogenetic tree was constructed with MEGA5 software (Tamura et al. 2011) using the complete coding sequences obtained together with the LMWGs genes identified in *T. urartu* by Luo et al. (2015), and the LMW-i sequences from *T. urartu* obtained in a previous study (Cuesta et al. 2015). Neighbour-joining cluster with all sequences analysed was generated using the Maximun Composite likelihood method (Tamura et al. 2004) and one bootstrap consensus from 1000 replicates was used (Felsenstein1985).

Results

The PCR detected up to six amplicons in each of the accessions evaluated (Fig. 1), and some were classified as LMW-i genes, which were not been included in this study (results not shown). In total, 21 sequences of LMWGs genes were sequenced: eight from wild einkorn, four from cultivated einkorn and nine from *T. urartu* (Table 1).

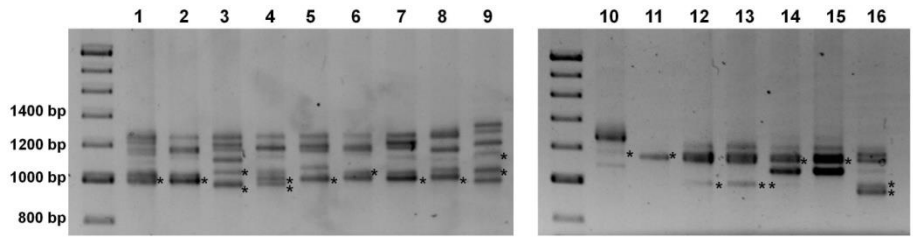


Figure 1. PCR amplification of LMWG genes from *T. boeoticum*, *T. monococcum*, and *T. urartu*. Lane as follows: 1, PI 427470; 2, PI 427505; 3, PI 427618; 4, PI 427749; 5, PI 573523; 6, PI 538524; 7, PI 355519; 8, PI 427959; 9, PI 428171; 10, PI 428183; 11, PI 428188; 12, PI428225; 13, PI428255; 14, PI 428327; 15, PI 428328; and 16, TRI 6734. Each asterisk identifies one amplicon sequenced.

The BLAST algorithm indicated that only 15 of these sequences were unique, being classified as LMW-m (13) or LMW-s (2) according to the first amino acid residue of the deduced mature protein, and 14 had not been previously described. The sizes of these novel alleles ranged from 864 to 930 bp for LMW-m genes and from 989 to 996 bp for LMW-s genes (Table 1). Ten, eight LMW-m and two LMW-s, out of these 15 alleles showed deduced truncated proteins and were classified as pseudogenes, with the main cause of this inactivation being the presence of one or more in-frame stop codons.

The relationships between the alleles obtained and the LMWGs genes sequenced in *T. urartu* by Luo et al. (2015), together with the LMW-i sequences obtained in our previous study (Cuesta et al. 2015), were analysed using a phenogram based on the maximum composite likelihood method (Fig. 2). Fourteen of the alleles detected were associated with some of the genes described by Luo et al. (2015), while the KR024661 sequence was not associated with any of those genes and appeared as clearly separated from the rest of the sequences (Fig. 2). Among the LMW-m, eight of the sequences were associated with the *TuA3-391* gene, five with the *TuA3-391*/*TuA3-*392 group and the other three with the *TuA3-373* genes. The main differences between the groups are two stop codons in the sequences (one in the repetitive domain and the other in the C-II domain) of the *TuA3-391*/*TuA3-392* group that are absent in the other group. Two of these sequences showed additional stop codons in the repetitive domain (KR024649, position 124–126, and KR024648/KR024656, position 259–261) that were also absent in the *TuA3-373* sequence group. Furthermore, one of these sequences present one stop codon localised to the C-I domain (KM0100189, position 391–393). However, the two

TuA3-373 pseudogenes contain a unique stop codon in the repetitive domain (position 202–204) and one insertion/deletion (InDel) of six nucleotides (position 154–159), which are not present in the sequences of the *TuA3-391*/*TuA3-392* group.

Table 1. LMWG sequences obtained in this study.

LMW type	Gene[a]	NCBI ID[b]	Accession	DNA size
T. monococcum ssp. *aegilopoides* (wild einkorn)				
LMW-m	*TuA3-391*	KR024646	PI 427470	888
		KR024647	PI 427505	888
		KR024648	PI 427618	864
		KR024649	PI 427618	864
		KR024650	PI 427749	885
		KR024652	PI 573523	885
		KR024653	PI 538524	891
	TuA3-400	**KR024651**	PI 427749	906
T. monococcum ssp. *monococcum* (einkorn)				
LMW-m	*TuA3-391*	KR024654	PI 355519	885
		KR024656	PI 428171	864
	TuA3-400	**KR024655**	PI 427959	885
LMW-s	*TuA3-460*	KR024657	PI 428171	996
T. urartu				
LMW-m	*TuA3-391*	KM010189	TRI 6734	885
	TuA3-397	KR024659	PI 428225	909
		KR024660	PI 428255	909
	TuA3-400	**KM010190**	TRI 6734	906
LMW-m/-i	New	**KR024661**	PI 428255	930
LMW-s	*TuA3-460*	KM010188	PI 428183	989
		KR024658	PI 428188	989
		KR024662	PI 428327	989
		KR024663	PI 428328	989

[a] according to Luo et al (2015).
[b] the sequences in bold presented active protein. The rest are pseudogenes.

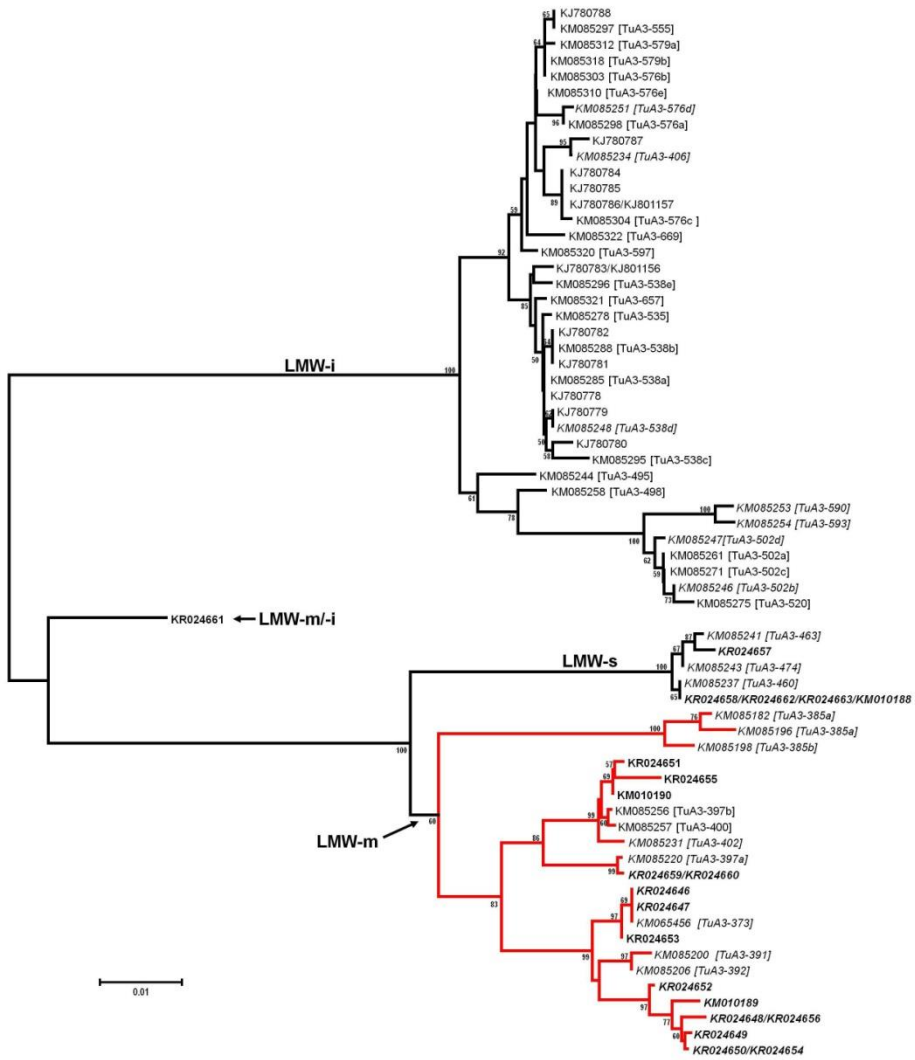


Figure 2. Neighbour-joining tree based on the maximum composite likelihood method of LMWG gene sequences detected in einkorn and *T. urartu* accessions (bold), together with the previous sequences described of *T. urartu* and common wheat. Numbers in nodes indicate bootstrap estimates from 1000 replications.

One of the pseudogenes detected in two *T. urartu* accessions (KR024659/KR024660) was present in the *TuA3-397* gene, which contains a unique stop codon derived from the change of C → T (position 724) inside the C-II domain.

For the LMW-s gene, two pseudogenes were found, one in an einkorn accession (KR024657) associated with the *TuA3-463* group, and the other in four accessions of *T. urartu* (KM010188/KR024658/KR024662/KR024663) inside the *TuA3-460* group, which

has six nucleotides less than in the repetitive domain. Although both pseudogenes present a stop codon at the beginning of the repetitive domain (position 145–147), the C-II domain had marked differences. The sequence of *T. urartu* presents one single nucleotide polymorphisms (SNP, C720 → -) that generate one frame-shift mutation. Additionally, the KR024657 sequence presents a double stop codon 30-bp upstream inside the C-II domain, which is absent in the other sequence, where this change in the frame-shift generates three stop codons in the C-III domain.

The remaining five alleles found encoded complete mature proteins with intact open reading frames (ORFs). Three are associated with *TuA3-397b*, *TuA3-400* and *TuA3-402*, which corroborates the findings of Luo et al. (2015) who detected active genes for *TuA3-397b* and *TuA3-400*. However, they established two separate groups for these alleles. The bootstrap values of the current dendrogram suggest that these alleles form a group together with the *TuA3-402*, clearly separate from *TuA3-397a*, where all of the variants found were inactive both in Luo et al. (2015) and in the current study.

Contrary to previous findings that showed only inactive variants for the *TuA3-391* gene, the KR024653 sequence detected in one accession of wild einkorn that appeared to be associated with the gene, and with *TuA3-373*, is active. The other active allele was KR024661, which is not associated with any of the genes described by Luo et al. (2015).

Protein analysis of LMW-m variants

The deduced proteins from active LMWGs genes isolated in the current study were compared with the *TuA3-397b* variant isolated from *T. urartu* by Luo et al. (2015), which was an active gene (Fig. 3). Compared with the reference sequences, no extra cysteine residues were found and the typical eight cysteine residues of the LMWGs were conserved in all of the sequences.

Amino acid deduced sequences were analysed for the presence of the four glutenin epitopes (Glt-17: PFSQQQQQ; Glt-156: PFSQQQQSPF; Glu-21: QSEQSQQPFQPQ; and Glu-5: QI/LQQPQQF) described by Vader et al. (2002) as stimulating T-cells in celiac disease. Only the Glt-17 epitope was found once in the repetitive domains of the four sequences (KR024651, KR024655, KR024661 and KM010190), whereas the KR024653 sequence from wild einkorn showed one SNP inside this target motif (PFSPQQQQ; Fig. 3).

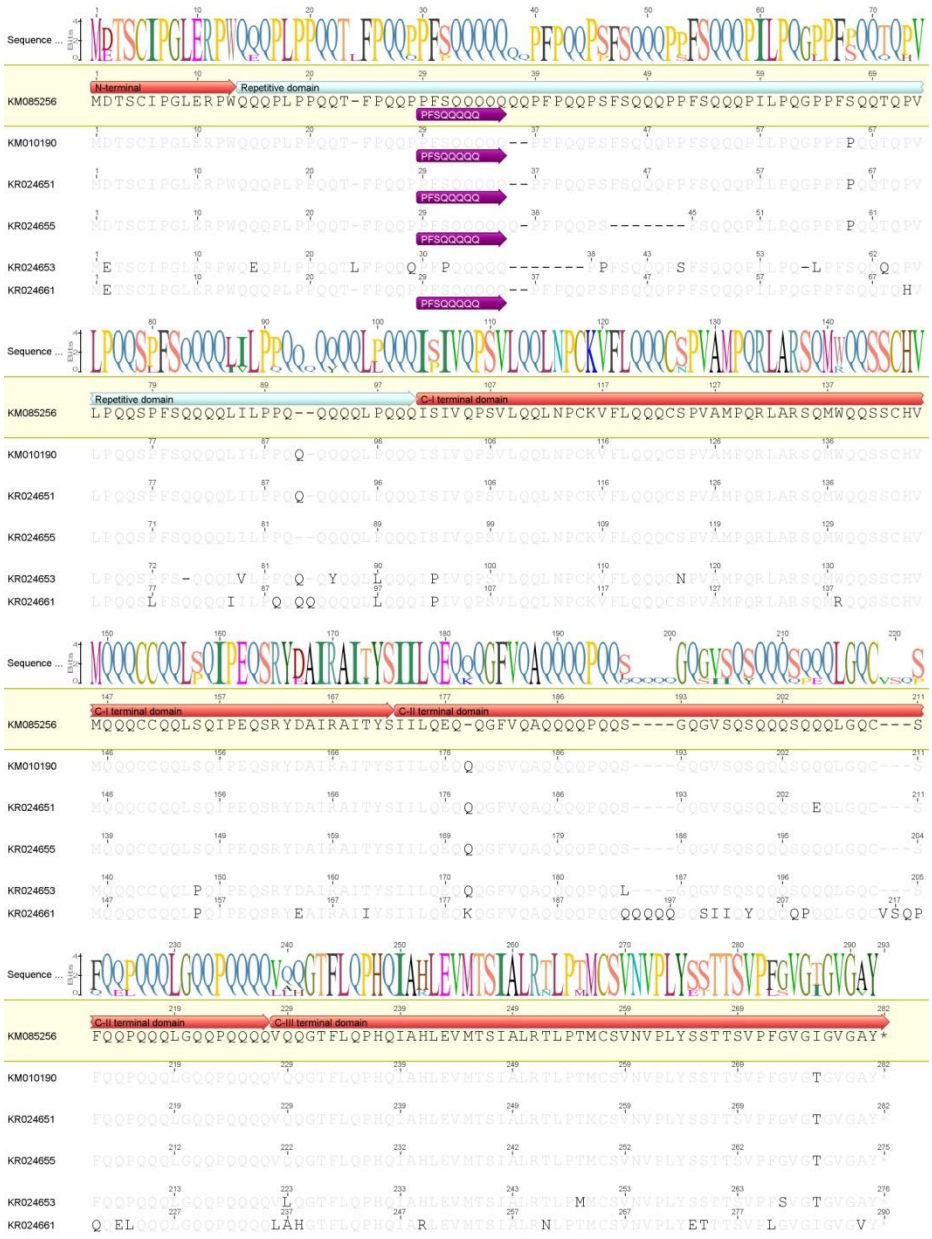


Figure 3. Alignments of the aminoacid sequences of einkorn and *T. urartu* with respect to the *TuA3-397b* (KM085266) identified as active LMW-m by Luo et al (2015), and identification of Glt-17 (PFSQQQQQ) epitope.

Two sequence groups (MDTSCIPGLERPW and METSCIPGLERPW) were detected based on the N-terminal domain. Three sequences (KR024651, KR024655 and

KM010190) were similar to KM085256, which was used as a reference, while the other two sequences (KR024653 and KR024661) showed an Asp2 → Glu change. Although several InDels were found within the repetitive domain, most of them were of one or two residues, with the exception of two with seven residues: FSQQQPP in KR024655 and QQPFPQQ in KR024653.

In the rest of the domains, the changes were mainly substitutions of amino acid residues, such as proline and glutamine, which could affect the LMWGs' structure (Fig. 3). These changes were different in each of the abovementioned groups. The MDTSCIPGLERPW subunits showed three changes with respect to the reference. All of them presented one additional glutamine at the beginning of the C-II domain and the Ile → Thr substitution at the end of the C-III domain (Fig. 3). Additionally, KR024651 has a change at the C-II domain (Gln205 → Glu).

The METSCIPGLERPW sequences showed clear differences between them and those of the other group. KR024653 has nine amino acid substitutions (three at the C-I, two at the C-II and four at the C-III domain) compared with the other group. Only two of these substitutions were common to other sequences of this group, in both cases being the Ser → Pro change within the C-I domain. The rest of the KR024661 sequence displayed some unusual features (Fig. 3) that justified the analysis of this sequence.

Identification of the corresponding subunits

The compositions of LMW-GSs in accessions with active genes were identified by SDS-PAGE (Fig. 4). There are two clear regions, which corresponded to B-subunits and C-subunits. The former contained 1-3 bands and the second was composed of several bands. For the KR024651 and KM010190 alleles, the size and weight of the deduced mature protein was 281 residues and 31.77 kDa, respectively. For KR024653 and KR024655 the sizes were 276 and 274 residues, being the weights 31.46 kDa and 30.95 kDa, respectively. The subunits in the SDS-PAGE gel were identified by calculating accurate M_r values using MALDI-TOF-MS (Fig. 4). For the KR024651, KR024653 and KM010190 sequences, slight differences between the M_r values determined by MS and the deduced mass were registered (Table 2). The differences could be the result of post-translational modification as reported by other authors (Laurière et al. 1996; An et al. 2006). The M_r values of active genes containing these subunits corresponded to those of C-subunits (Fig. 4).

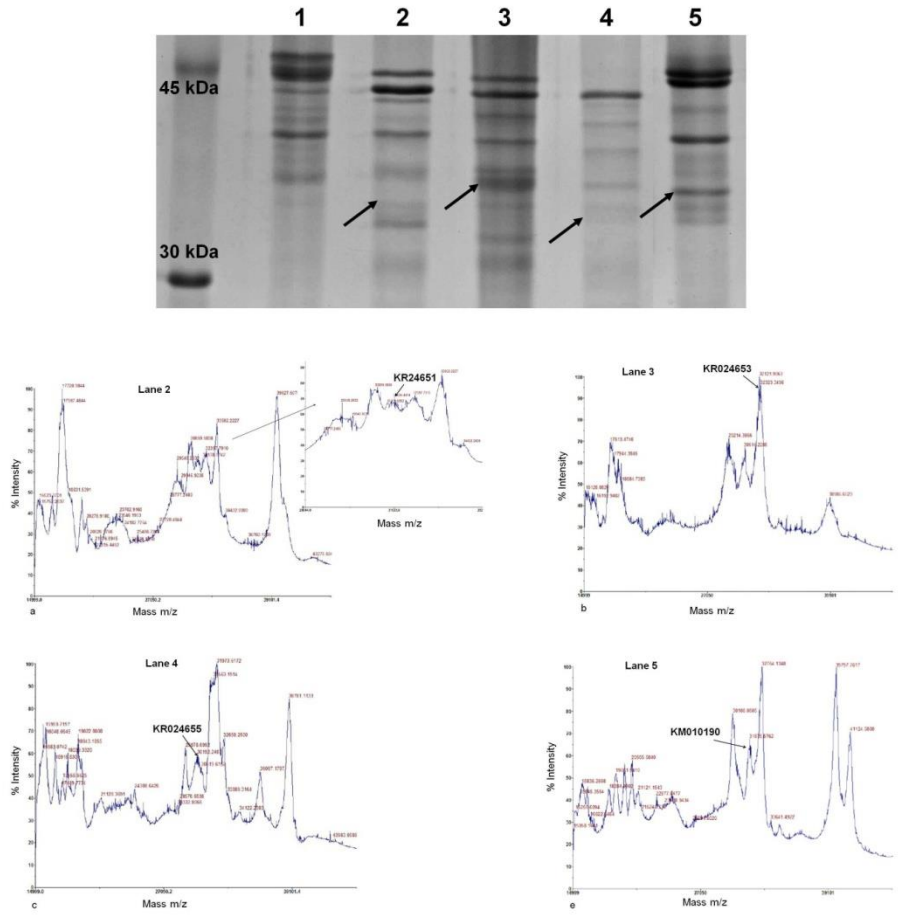


Figure 4. SDS-PAGE separation and MALDI-TOF-MS of the LMWG subunits evaluated in this study. Lanes as follows: 1, cv. Chinese Spring; 2, PI 427749; 3, PI 538524; 4, PI 427959; 5, TRI 6734. Arrows indicate the subunits sequenced.

Protein analysis of KR024661

The alignment of the KR024661 sequence showed two different regions in the amino acid sequence: one between the N-terminal domain and the beginning of the C-II domain, and the other one between the ends of the C-II domain and the protein (Fig. 3). The first region presents a high similarity with the type-m subunits, including the presence of an N-terminal domain that begins with the Met residue, whereas the latter region is better aligned with the type-i subunits, including the termination sequence VGIGVGVY, which was identical to those of the LMW-i genes (Ikeda et al. 2002). In

fact, the cluster analysis showed the intermediate position of this sequence between the groups (Fig. 2).

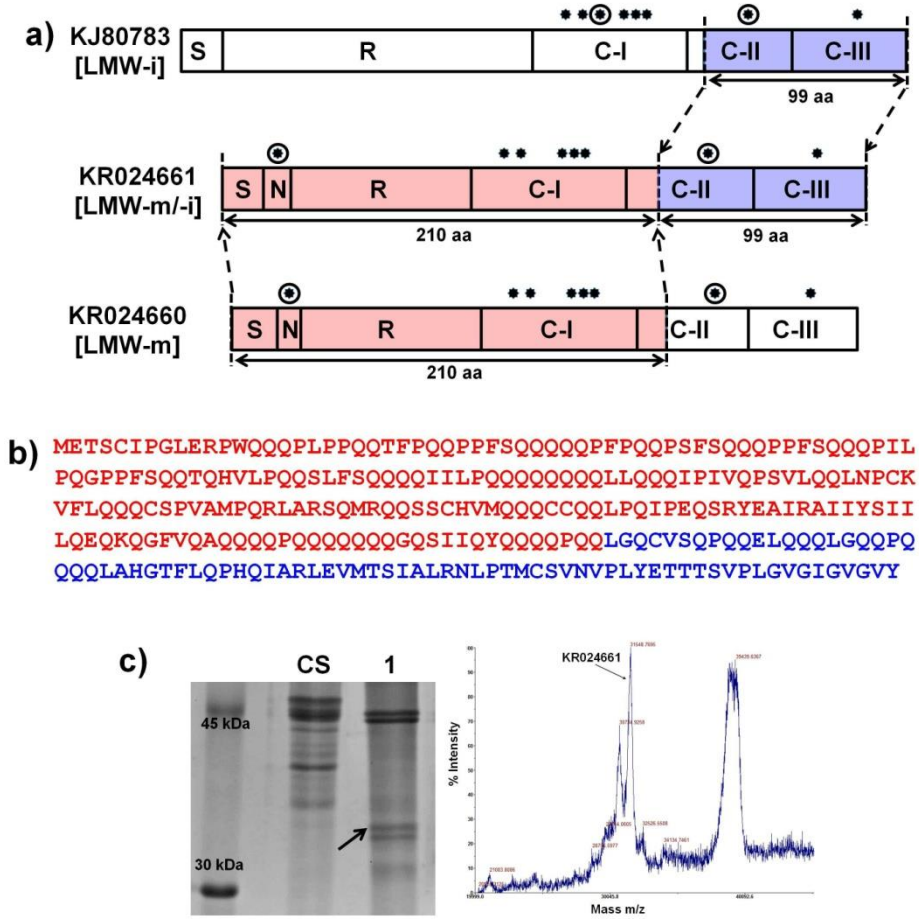


Figure 5. **a)** Diagram representing the LMW-m/LMW-i chimeric gene (KR024661) and their putative ancestors (KJ780783 LMW-i gene and KR024660 LMW-m gene); **b**) amino acid deduced sequence; and **c**) SDS-PAGE and MALDI-TOF-MS profile of the PI 428255 accession. Blue and red frame represent LMW-i and LMW-m genes fraction, respectively, involved in recombination. S: Signal peptide; N: N-terminal domain; R: repetitive domain; C-I, C-II and C-III: C-terminal domains. The cysteine residues are indicated as asterisks. The encircled asterisks represent cysteine residues involved in inter-molecular disulfide bonds.

In a further analysis, this subunit was compared with those found in the same accession. The alignment showed that the first region was 99.4% similar to KR024660 (a LMW-m pseudogene), while the second region displayed a similarity of 89.5% with the

KJ780783 sequence, a LMW-i gene sequenced in our previous study (Cuesta et al. 2015). All of these data suggested that this new sequence (KR024661) was a chimeric gene, whose deduced protein was 289 residues with an M_r of 33.11 kDa, and the 1–210 residues had characteristics of an LMW-m protein, while residues 211–289 had characteristics of an LMW-i protein (Fig. 5a and b). The identification of this chimeric protein was also performed by MALDI-TOF-MS, showing a M_r of 31.55 kDa (Fig. 5c and Table 2).

Discussion

The LMW-m and LMW-s genes have been reported for the *Glu-B3* and *Glu-D3* loci. On the contrary, the *Glu-A3* locus had been considered mainly a source of LMW-i genes. However, several studies (Lee et al. 1999a; Dong et al. 2010; Zhang et al. 2013) detected the presence of LMW-m genes at the *Glu-A3* locus, although the sequences detected were generally catalogued as pseudogenes. More recently, Luo et al. (2015) found LMW-s pseudogenes in *T. urartu*. This suggested that the three subunit types could be present in the species donor of the wheat A genome, which made it possible to increase the variability of these LMWGs in wheat by the introgression of these genes from diploid wheat, such as *T. urartu*.

Although previous classifications have been performed (Dong et al. 2010; Wang et al. 2010; Zhang et al. 2013), more recently Luo et al. (2015) suggested the presence of eight LMWGs genes (four for LMW-m, one for LMW-s, and three for LMW-i) in *T. urartu*. All of these genes, with the exception of the LMW-s, have corresponding LMWGs genes described in common wheat by Zhang et al. (2013). In our previous study (Cuesta et al. 2015), the variability of the LMW-i genes in *T. urartu* was evaluated, and a high degree of variation was detected for these genes. In the current study, the variation detected for the LMW-m genes was also high, with 15 novel alleles of the LMW-m (13) and LMW-s (2) genes identified in three species of diploid wheat, including *T. urartu*. Ten of these 15 alleles were pseudogenes, which are common in cereal prolamins due to the great number of glutamine residues. Glutamine is codified by CAA and CAG, which can each be changed to the stop codons TAA and TAG, respectively, by single nucleotide mutations. This data corroborates that of earlier studies (Han et al. 2011; Luo et al. 2015; Zhang et al. 2013), where numerous pseudogenes of LMWGs genes were reported.

In the five sequences with intact ORFs, SNPs and InDels were detected that could result in functional differences in the encoded proteins (Ma et al. 2006). Particularly,

those changes detected in proline and glutamine residues could alter protein structures and, therefore, affect gluten strength (Tanaka et al. 2005). However, no changes were found in the cysteine residues, which were all conserved, without novel cysteine residues in the protein sequence. This is important because cysteine residues are needed in the formation of inter- and intra-molecular disulfide bonds (Shewry et al. 1995; D'Ovidio and Masci 2004), which affects the protein structure and its capacity to stabilize links with other adjacent proteins.

Another important feature of these proteins is their relationship with celiac disease because, although the main epitopes causing this disease are present in gliadins, some research has indicated that glutenin (HMWGs and LMWGs) has some epitopes (Vader et al. 2002). The number of epitopes found in each subunit was less than that detected for the LMW-i subunits previously evaluated in *T. urartu* (Cuesta et al. 2015), and one sequence, KR024653, had no detectable epitopes. This suggests that the LMW-m subunit could be less reactive than the LMW-i subunits in celiac disease, although, due to their relatively low level of influence on this disease, their final effects would be conditioned by the presence of highly reactive gliadins.

According to the data of Luo et al. (2015), eight out of the 13 LMW-m alleles were associated with the *TuA3-391* gene. The number of allelic variants exceeded that found by Zhang et al. (2013) in common wheat (five variants), or by Luo et al. (2015) in *T. urartu* (three variants). Most of the allelic variants in which the *TuA3-391* gene was sequenced were from wild einkorn, and one of them had an intact ORF.

The other five LMW-m alleles appeared to be associated with the *TuA3-400/TuA3-397* group. In the current study, the number of variants (four) found for this group was identical to that found in *T. urartu* by Luo et al. (2015), but less than that found in common wheat, which had seven variants (Zhang et al. 2013). Luo et al. (2015) hypothesized that the *A3-400* from common wheat and the *TuA3-397* from *T. urartu* should be homologs and that the former was derived from a duplication of the latter in *T. urartu*. In fact, no variant of the *TuA3-397* gene was detected in wild or cultivated einkorn. This suggests that the duplication event could only have occurred in *T. urartu* and not in the other A-genome diploid species.

Until now, studies performed with common wheat have suggested that the LMWGs synthesized in the A genome are exclusively LMW-i and LMW-m (Zhang et al. 2013). However, Luo et al. (2015) identified for the first time a LMW-s gene in the *Glu-A3* locus of *T. urartu*, which was named *TuA3-460* and had three variants. In the current

study, two LMW-s sequences were detected and analysed, although unfortunately both were pseudogenes. One of these alleles was detected in *T. urartu* and shared 100% identity with the *TuA3-460* variant reported by Luo et al. (2015), and it was identified as being widespread among the accessions in *T. urartu*. Another novel variant was identified from cultivated einkorn. In this respect, Luo et al. (2015) suggested that the *TuA3-460* gene did not share the same evolutionary process as other LMW-s genes of common wheat from the primitive LMW-m or that they originate from different LMW-m genes. The discovery of one variant of this LMW-s gene in cultivated einkorn suggests that this gene could have emerged in the common ancestor of the A-genome diploid species.

The presence of chimeras has been reported among the different storage protein genes (Nagy et al. 2005; Li et al. 2008b, c; Qin et al. 2015). This process could be related to an illegitimate recombination event, where the repetitive structure of these genes, together with the complex nature of the *Glu-3* loci, is key. This genetic mechanism has been involved in the origin of these genes and is important for creating allelic variations that can be used for wheat quality improvement (Yuan et al. 2011). Li et al. (2008b) detected a chimeric gene derived from recombination between LMW-i and LMW-m genes in two *Aegilops* species, *Ae. juvenalis* and *Ae. kotschyi*. This gene showed the typical LMW-i gene structure at the beginning of the sequence, and was similar to LMW-m in its latter domains. Our data showed one novel chimeric gene combination of LMW-i and LMW-m types in one accession of *T. urartu*, but in this case the structure was the reverse of that described by Li et al. (2008b). The beginning of the sequence was typical of the LMW-m genes, while the end of the sequence was similar to LMW-i. The putative donor genes were also detected in this accession. The combination of LMW-m and LMW-i features in the chimeric gene could lead to differences in dough quality and may be involved in the development of new varieties that are adapted to new uses.

Acknowledgements

This research was supported by Grant AGL2014-52445-R from the Spanish Ministry of Economy and Competitiveness, co-financed by the European Regional Development Fund (FEDER) from the European Union; and Grant P11-AGR-7920 from the Regional Government of Andalusia (Southern Spain). The first author is grateful to the Spanish Ministry of Economy and Competitiveness (FPI programme) and European Social Fund for a predoctoral fellowship. We thank to Dr. E. Chicano (Proteomic Unit, IMIBIC, Córdoba, Spain) for the MALDI-TOF-MS analysis.

DISCUSIÓN GENERAL Y CONCLUSIONES

En la actualidad existe una necesidad de ampliar la base genética de los cultivos. Por esta razón, es fundamental la exploración de la diversidad genética almacenada en los Bancos de Germoplasma con el fin de buscar nuevos alelos que podrían ser útiles en la mejora de caracteres de interés del trigo moderno. En base a esto, en esta Tesis Doctoral se han evaluado colecciones de germoplasma pertenecientes a especies diploides del género *Aegilops* y *Triticum*.

Considerando globalmente el trabajo, se ha caracterizado la variabilidad de los genes *Pin*, *Gsp-1* y de LMWGs en especies diploides del género *Aegilops*. En el caso de las especies del género *Triticum*, dado que los genes *Pin* habían sido previamente analizado en otro trabajo previo del grupo (Guzmán et al. 2012), se procedió directamente a la caracterización molecular de los genes de LMWGs, así como la identificación de las correspondientes subunidades en SDS-PAGE mediante MALDI-TOF-MS.

Para los genes *Pin* y *Gsp-1*, un alto polimorfismo fue encontrado, con la detección de numerosos alelos nuevos. Los alelos revelaron nuevas mutaciones que podrían afectar a la dureza del endospermo. La más significativa fue un cambio que afectaba al codón de inicio de PINA que puede generar una reducción en la expresión de la proteína, lo cual podría resultar en la expresión de una dureza intermedia entre *soft* y *hard*. Este tipo de cambio es raro en células eucariotas y ha sido descrito solo en algunas ocasiones, pero nunca en los genes *Pin*. Para el gen *Gsp-1*, fue identificada una mutación que podría alterar la estabilidad de la proteína. La eliminación de la primera cisteína de la proteína madura podría resultar en incrementos de la dureza como ha sido previamente demostrado para las PINs.

La caracterización de los genes de LMWG en especies de *Aegilops* y *Triticum* reveló igualmente numerosos alelos nuevos y algunos genes no descritos hasta el momento. El más relevante fue la detección de una variante quimérica en una accesión de *T. urartu* derivada de la recombinación entre un gen de LMW-m y un gen de LMW-i (LMW-m/LMW-i), que podría tener un efecto sobre la calidad de la masa diferente al descrito hasta el momento para cada una de las subunidades por separado. En general, mediante el análisis MALDI-TOF-MS se pudo establecer una relación entre los genes caracterizados y la banda correspondiente a la subunidad de proteína codificada detectada

mediante SDS-PAGE para las especies de *Triticum*. Además, el análisis de los epítopos reactivos para la enfermedad celíaca sugirío que las subunidades de LMW-i detectadas en *T. urartu* fueron las más reactivas, mientras que las LMW-s de la sección *Sitopsis* de *Aegilops* fueron las menos reactivas. Las subunidades de LMW-m presentarían un número intermedio de epítopos en todas las especies, aunque en un gen identificado en el genoma S de *Aegilops*, el cual no está presente en el genoma B del trigo, no se presentó ningún epítopo reactivo.

Las relaciones filogenéticas de las secuencias obtenidas con los genes del trigo común demostró que los genes *Gsp-1* son poco eficaces para estudios filogenéticos, debido a su reducido tamaño. Por el contario, el estudio filogenético de las secuencias de los genes de LMWG reveló nuevos datos sobre la evolución y composición de esta familia multigénica. Del mismo modo, los genes de LMWG permitieron estudiar las relaciones filogenéticas entre las especies estudiadas, corroborando la hipótesis de que el genoma B de los trigos poliploides podría tener un origen polifilético.

A partir del trabajo realizado se han obtenido las siguientes conclusiones:

1. Las especies diploides de *Aegilops* se presentan como una buena fuente de variabilidad para los genes *Pin* y *Gsp-1* que podrían generar cambios en la dureza.
2. La variabilidad de los genes de LMWG detectada en la sección *Sitopsis* de *Aegilops*, muestra que los alelos encontrados podrían presentar una ventaja en el desarrollo de nuevos cultivares de trigo que exhibieran nuevas propiedades en su harina, incluido un potencial uso en la elaboración de productos aptos para celíacos.
3. Las especies diploides de *Triticum* son una buena fuente de variabilidad, siendo confirmado este hecho con la detección, incluso, de una variante quimérica LMW-m/LMW-i que podría tener un efecto único sobre la calidad de la masa de harina.
4. Los genes *Gsp-1*, no son útiles para el estudio de relaciones filogenéticas, mientras que los genes de LMWG sí que lo serían. Además el análisis filogenético puede ser una buena herramienta en el estudio de la evolución y composición de la familia multigénica de los genes de LMWG.
5. Como consecuencia de todo lo anterior, es esencial el mantenimiento de estas especies con el fin de salvaguardar su diversidad para que puedan ser utilizadas como recursos genéticos en programas de mejora del trigo actual.

BIBLIOGRAFÍA

Aizat WM, Preuss JM, Johnson AAT, Tester MA, Schultz CJ (2011) Investigation of a His-rich arabinogalactan-protein for micronutrient biofortification of cereal grain. Physiol Plant 143:271-286

Alvarez JB, Moral A, Martín LM (2006) Polymorphism and genetic diversity for the seed storage proteins in Spanish cultivated einkorn wheat (*Triticum monococcum* L. ssp. *monococcum*). Genet Resour Crop Evol 53:1061-1067

Alvarez JB, Ballesteros J, Sillero JA, Martín LM (1992) Tritordeum: a new crop of potential importance in the food industry. Hereditas 116: 193-197

Alvarez JB, Gutiérrez MV, Guzmán C, Martín LM (2013) Molecular characterisation of the amino-and carboxyl-domains in different *Glu-A1x* alleles of *Triticum urartu* Thum. ex Gandil. Theor Appl Genet 126:1703-1711

Alvarez JB, Martín A, Martín LM (2001) Variation in the high molecular weight glutenin subunits coded at the *Glu-H^{ch}1* locus in *Hordeum chilense*. Theor Appl Genet 102:134-137

Alvarez JB, Moral A, Martín LM (2006) Polymorphism and genetic diversity for the seed storage proteins in Spanish cultivated einkorn wheat (*Triticum monococcum* L. ssp. *monococcum*). Genet Resour Crop Evol 53:1061-1067

An X, Zhang Q, Yan Y, Li Q, Zhang Y, Wang A, Pei Y, Tian J, Wang H, Hsam SLK, Zeller FJ (2006) Cloning and molecular characterization of three novel LMW-i glutenin subunit genes from cultivated einkorn (*Triticum monococcum* L.). Theor Appl Genet 113:383-395

Barak S, Mudgil D, Khatkar B (2015) Biochemical and functional properties of wheat gliadins: A Review. Crit Rev Food Sci Nutr 55:357-368

Baum BR, Bailey LG (2004) The origin of the A genome donor of wheats (*Triticum*: Poaceae) – a perspective based on the sequence variation of the 5S DNA gene units. Genet Resour Crop Evol 51:183-196

Bell M, Fischer R, Byerlee D, Sayre K (1995) Genetic and agronomic contributions to yield gains: A case study for wheat. Field Crops Res 44:55-65

Bettge A, Morris CF (2000) Relationships among grain hardness, pentosan fractions, and end-use quality of wheat. Cereal Chem 77:241-247

Bhave M, Morris C (2008a) Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. Plant Mol Biol 66:205-219

Bhave M, Morris C.F (2008b) Molecular genetics of puroindolines and related genes: regulation of expression, membrane binding properties and applications. Plant Mol Biol 66:221-231

Blochet J-E, Chevalier C, Forest E, Pebay-Peyroula E, Gautier M-F, Joudrier P, Pézolet M, Marion D (1993) Complete amino acid sequence of puroindoline, a new basic and cystine-rich protein with a unique tryptophan-rich domain, isolated from wheat endosperm by Triton X-114 phase partitioning. FEBS Letters 329:336-340

Bushuk W, Zillman R (1978) Wheat cultivar identification by gliadin electrophoregrams. I. Apparatus, method and nomenclature. Can J Plant Sci 58:505-515

Caballero L, Martín MA, Alvarez JB (2009) Genetic diversity for seed storage proteins in Lebanon and Turkey populations of wild diploid wheat (*Triticum urartu* Thum. ex Gandil.). Genet Resour Crop Evol 56:1117-1124

Caballero L, Martín MA, Alvarez JB (2008) Allelic variation for the high- and low-molecular-weight glutenin subunits in wild diploid wheat (*Triticum urartu*) and its comparison with durum wheats. Aust J Agric Res 59:906-910

Campbell KG, Bergman CJ, Gualberto DG, Anderson JA, Giroux MJ, Hareland G, Fulcher RG, Sorrells ME, Finney PL (1999) Quantitative trait loci associated with kernel traits in a soft × hard wheat cross. Crop Sci 39:1184-1195

Cassidy BG, Dvorak J (1991) Molecular characterization of a low-molecular-weight glutenin cDNA clone from *Triticum durum*. Theor Appl Genet 81:653-660

Cassidy BG, Dvorak J, Anderson OD (1998) The wheat low-molecular-weight glutenin genes: characterization of six new genes and progress in understanding gene family structure. Theor Appl Genet 96:743-750

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier M.F, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). Plant Cell 17:1033-1045

Chao S, Sharp P, Worland A, Warham E, Koebner R, Gale M (1989) RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. Theor Appl Genet 78:495-504

Chen F, Zhao F, Liu R, Xia G (2011) Functional properties of two low-molecular-weight glutenin subunits carrying additional cysteine residues from hybrid introgression line II-12 derived from *Triticum aestivum* and *Agropyron elongatum*. Food Chem 127:1773-1776

Chen M, Wilkinson M, Tosi P, He G, Shewry P (2005) Novel puroindoline and grain softness protein alleles in *Aegilops* species with the C, D, S, M and U genomes. Theor Appl Genet 111:1159-1166

Chen P, Li R, Zhou R, Zhiguo E, He G (2010) Cloning and characterization of novel low molecular weight glutenin subunit genes from two *Aegilops* species with the C and D genomes. Genet Resour Crop Evol 57:881-890

Christensen A, Lyznik A, Mohammed S, Elowsky CG, Elo A, Yule R, Mackenzie SA (2005) Dual-domain, dual-targeting organellar protein presequences in *Arabidopsis* can use non-AUG start codons. Plant Cell 17:2805-2816

Cloutier S, Rampitsch C, Penner GA, Lukow OM (2001) Cloning and expression of a LMW-i glutenin gene. J Cereal Sci 33:143-154

Corona V, Gazza L, Boggini G, Pogna NE (2001) Variation in friabilin composition as determined by A-PAGE fractionation and PCR amplification, and its relationship to grain hardness in bread wheat. J Cereal Sci 34:243-250

Cubero JI (2003) Introducción a la Mejora Genética Vegetal. 2ª Ed. Mundi-Prensa. Madrid

Cuesta S, Guzmán C, Alvarez JB (2015) Molecular characterization of novel LMW-i glutenin subunit genes from *Triticum urartu* Thum. ex Gandil. Theor Appl Genet (in press, DOI: 10.1007/s00122-015-2574-1)

Cuesta S, Guzmán C, Alvarez JB (2013) Allelic diversity and molecular characterization of *Puroindoline* genes in five diploid species of the *Aegilops* genus. J Exp Bot 64:5133-5143

Darlington HF, Rouster J, Hoffmann L, Halford NG, Shewry PR, Simpson DJ (2001) Identification and molecular characterisation of hordoindolines from barley grain. Plant Mol Biol 47:785-794

Dong L, Zhang X, Liu D, Fan H, Sun J, Zhang Z, Qin H, Li B, Hao S, Li Z, Wang D, Zhang A, Ling HQ (2010) New insights into the organization, recombination, expression and functional mechanism of low molecular weight glutenin subunit genes in bread wheat. PloS One 5:e13548

Douliez JP, Michon T, Elmorjani K, Marion D (2000) Structure, biological and technological functions of lipid transfer proteins and indulines, the major lipid binding proteins from cereal kernels. J Cereal Sci 32:1-20

D'Ovidio R, Masci S (2004) The low-molecular-weight glutenin subunits of wheat gluten. J Cereal Sci 39:321-339

Dvorak J, Akhunov ED (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. Genetics 171:323-332

Dvorak J, Luo M-C, Yang Z-L (1998) Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics 148:423

Dvorak J, Terlizzi Pd, Zhang H-B, Resta P (1993) The evolution of polyploid wheats: identification of the A genome donor species. Genome 36:21-31

Egidi E, Sestili F, Janni M, D'Ovidio R, Lafiandra D, Ceriotti A, Vensel WH, Kasarda DD, Masci S (2014) An asparagine residue at the N-terminus affects the maturation process of low molecular weight glutenin subunits of wheat endosperm. BMC Plant Biol 14:64

Elmorjani K, Geneix N, Dalgalarrondo M, Branlard G, Marion D (2013) Wheat grain softness protein (*Gsp1*) is a puroindoline-like protein that displays a specific post-translational maturation and does not interact with lipids. J Cereal Sci 58:117-122

Esquinas-Alcázar J (2005) Protecting crop genetic diversity for food security: political, ethical and technical challenges. Nat Rev Genet 6:946-953

FAO. (2001). International Treaty for Plant Genetic Resources for Food and Agriculture. Rome, Italy: Food and Agriculture Organization. [ftp://ext-ftp.fao.org/ag/cgrfa/it/ITPGRe.pdf]

FAO (2013) http://faostat.fao.org/

Feiz L, Beecher BS, Martin JM, Giroux MJ (2009) In planta mutagenesis determines the functional regions of the wheat puroindoline proteins. Genetics 183:853-860

Felsenstein J (1985) Confidence-limits on phylogenies - an approach using the bootstrap. Evolution 39:783-791

Gautier M-F, Aleman M-E, Guirao A, Marion D, Joudrier P (1994) *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol Biol 25:43-57

Gautier MF, Cosson P, Guirao A, Alary R, Joudrier P (2000) Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid *Triticum* species. Plant Sci 153:81-91

Gazza L, Conti S, Taddei F, Pogna N (2006) Molecular characterization of puroindolines and their encoding genes in *Aegilops ventricosa*. Mol Breed 17:191-200

Gedye KR, Morris CF, Bettge AD (2004) Determination and evaluation of the sequence and textural effects of the puroindoline a and puroindoline b genes in a population of synthetic hexaploid wheat. Theor Appl Genet 109:1597-1603

Gepts P (1990) Genetic diversity of seed storage proteins in plants. En: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding and genetic resources. 1 edn. Slinauer Associattes, Sunderland, Massachussetts, pp 64-82

Giroux MJ, Morris CF (1997) A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theor Appl Genet 95:857-864

Giroux MJ, Morris CF. (1998) Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. Proc Natl Acad Sci USA 95:6262-6266

Giroux MJ, Talbert L, Habernicht DK, Lanning S, Hemphill A, Martin JM (2000) Association of puroindoline sequence type and grain hardness in hard red spring wheat. Crop Sci 40:370-374

Gollan P, Smith K., Bhave M (2007) *Gsp-1* genes comprise a multigene family in wheat that exhibits a unique combination of sequence diversity yet conservation. J Cereal Sci 45:184-198

Gordon K, Fütterer J, Hohn T (1992) Efficient initiation of translation at non-AUG triplets in plant cells. Plant J 2:809-813

Guzmán C, Alvarez JB (2015) Wheat waxy proteins: polymorphism, molecular characterization and effects on starch properties. Theor Appl Genet (in press, DOI 10.1007/s00122-015-2595-9)

Guzman C, Caballero L, Martín LM, Alvarez JB (2012) Waxy genes from spelt wheat: new alleles for modern wheat breeding and new phylogenetic inferences about the origin of this species. Ann Bot 110:1161-1171

Guzmán C, Caballero L, Martín MA, Alvarez JB (2011) Molecular characterization and diversity of the *Pina* and *Pinb* genes in cultivated and wild diploid wheat. Mol Breeding 30:1-10

Haider N (2013) The origin of the B-genome of bread wheat (*Triticum aestivum* L.). Russ J Genet 49:263-274

Hammer K (2003) A paradigm shift in the discipline of plant genetic resources. Genet Resour Crop Evol 50:3-10

Hammer K, Heller J, Engels J (2001) Monographs on underutilized and neglected crops. Genet Resour Crop Evol 48:3-5

Han C, Yan ZH, Dai SF, Liu DC, Wei YM, Zheng YL, Lan XJ, Peng YY (2011) Molecular characterization of LMW glutenin genes from *Taeniatherum* Nevski. Genet Resour Crop Evol 58:1029-1039

Harberd N, Bartels D, Thompson R (1986) DNA restriction-fragment variation in the gene family encoding high molecular weight (HMW) glutenin subunits of wheat. Biochem Genet 24:579-596

Harlan J (1992) Crop and man. American Society Agronomy, Crop Science Society of America. Madison, Wisconsin

Harlan JR (1981) The early history of wheat: earliest traces to the sack of Rome. En: Evans LT, WJ Peacok (eds) Wheat science-today and tomorrow. Cambridge: Camb Univ Press, pp 1:19

Harlan JR, deWet JMJ (1971) Toward a racional classification of cultivated plants. Taxon, 20:509-517

Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glemin S (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 24:1506-1517

Huang S, Sirikhachornkit A, Su X, Faris J, Gill B, Haselkorn R, Gornicki P (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. Proc Natl Acad Sci USA 99:8133-8138

Huang X-Q, Cloutier S (2008) Molecular characterization and genomic organization of low molecular weight glutenin subunit genes at the *Glu-3* loci in hexaploid wheat (*Triticum aestivum*). Theor Appl Genet 116:953-966

Huang Z, Long H, Jiang Q, Wei Y, Yan Z, Zheng Y (2010) Molecular characterization of novel low-molecular-weight glutenin genes in *Aegilops longissima*. J Appl Genet 51:9-18

Ikeda TM, Nagamine T, Fukuoka H, Yano H (2002) Identification of new low-molecular-weight glutenin subunit genes in wheat. Theor Appl Genet 104:680-687

Jackson E, Holt L, Payne P (1983) Characterisation of high molecular weight gliadin and low-molecular-weight glutenin subunits of wheat endosperm by two-dimensional electrophoresis and the chromosomal localisation of their controlling genes. Theor Appl Genet 66:29-37

Jauhar PP (1993) Alien gene transfer and genetic enrichment of bread wheat. En: Damania AB (ed.) Biodiversity and wheat improvement. ICARDA-A Wiley Sayce Publication. pp 103-119

Jiang C, Pei Y, Zhang Y, Li X, Yao D, Yan Y, Ma W, Hsam SLK, Zeller FJ (2008) Molecular cloning and characterization of four novel LMW glutenin subunit genes from *Aegilops longissima*, *Triticum dicoccoides* and *T. zhukovskyi*. Hereditas 145:92-98

Johal J, Gianibelli M, Rahman S, Morell M, Gale K (2004) Characterization of low-molecular-weight glutenin genes in *Aegilops tauschii*. Theor Appl Genet 109:1028-1040

Johnson BL (1975) Identification of the apparent B-genome donor of wheat. Can J Genet Cytol 17:21-39

Johnson BL, Dhaliwal HS (1976) Reproductive isolation of *Triticum boeoticum* and *T. urartu* and the origin of the tetraploid wheats. Am J Bot 63:1088-1094

Kilian B, Özkan H, Deusch O, Effgen S, Brandolini A, Kohl J, Martin W, Salamini F (2007) Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. Mol Biol Evol 24:217-227

Kimber G, Sears ER (1987) Evolution in the genus *Triticum* and the origin of cultivated wheat. En: Wheat and Wheat Improvement, 2nd Ed (Heyne EG, Ed.). American Society of Agronomy, Madison, WI. pp 154-164

Kobayashi Y, Dokiya Y, Kumazawa Y, Sugita M (2002) Non-AUG translation initiation of mRNA encoding plastid-targeted phage-type RNA polymerase in *Nicotina sylvestris.* Biochem Bioph Res Co 299**:**57-61

Kole C (2011) Wild Crop Relatives: Genomic and Breeding Resources: Cereals, vol 1. Springer Science & Business Media

Krishnamurthy K, Balconi C, Sherwood JE, Giroux MJ (2001) Wheat puroindolines enhance fungal disease resistance in transgenic rice. Mol Plant Microbe In 14:1255-1260

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680-685

Lafiandra D, Benedettelli S, Margiotta B, Spagnoletti-Zeuli P, Porceddu E (1990) Seed storage-proteins and wheat genetic resources. En: Srivastaca JP, A.B Damania (eds) Wheat genetic resources: meeting diverse needs. Aleppo, Syria, pp 73-87

Lafiandra D, Riccardi G, Shewry PR (2014) Improving cereal grain carbohydrates for diet and health. J Cereal Sci 59:312-326

Lauriere M, Bouchez I, Doyen C, Eynard L (1996) Identification of glycosylated forms of wheat storage proteins using two-dimensional electrophoresis and blotting. Electrophoresis 17:497-501

Lee YK, Bekes F, Gupta RB, Appels R, Morell MK (1999b) The low-molecular-weight glutenin subunt proteins of primitive wheats. I. Variation in A-genome species. Theor Appl Genet 98:119-125

Lee YK, Ciaffi M, Appels R, Morell MK (1999a) The low-molecular-weight glutenin subunt proteins of primitive wheats. II. The genes from A-genome species. Theor Appl Genet 98:126-134

Li W, Huang L, Gill BS (2008a) Recurrent deletions of puroindoline genes at the grain *Hardness* locus in four independent lineages of polyploid wheat. Plant Physiol 146**:**200-212

Li XH, Ma WJ, Gao LY, Zhang YZ, Wang AL, Ji KM, Wang K, Appels R, Yan YM (2008b) A novel chimeric low-molecular-weight glutenin subunit gene from the

wild relatives of wheat *Aegilops kotschyi* and *Ae. juvenalis*: Evolution at the *Glu-3* loci. Genetics 180:93-101

Li XH, Wang K, Wang SL, Gao LY, Xie XX, Hsam SLK, Zeller FJ, Yan YM (2010) Molecular characterization and comparative transcriptional analysis of LMW-m-type genes from wheat (*Triticum aestivum* L.) and *Aegilops* species. Theor Appl Genet 121:845-856

Li ZX, Zhang XQ, Zhang HG, Cao SH, Wang DW, Hao ST, Li LH, Li HJ, Wang XP (2008c) Isolation and characterization of a novel variant of HMW glutenin subunit gene from the S^t genome of *Pseudoroegneria stipifolia*. J Cereal Sci 47:429-437

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452

Lillemo M, Cosimo Simeone M, Morris C (2002) Analysis of puroindoline a and b sequences from *Triticum aestivum* cv. "Penawawa" and related diploid taxa. Euphytica 126:321-331

Lillemo M, Morris CF (2000) A leucine to proline mutation in puroindoline b is frequently present in hard wheats from Northern Europe. Theor Appli Genet 100:1100-1107

Lin-Hai W, Min Z, Hui-Ling L, Zhonghu H, Xian-Chun X (2010) Cloning and phylogenetic analysis of low-molecular-weight glutenin subunit genes at *Glu-B3* locus in common wheat relative species. Hereditas (Beijing) 32:613-624

Long H, Huang Z, Wei Y-M, Yan Z-H, Ma Z-C, Zheng Y-L (2008) Length variation of i-type low-molecular-weight glutenin subunit genes in diploid wheats. Russ J Genet 44:429-435

Luo G, Zhang X, Zhang Y, Yang W, Li Y, Sun J, Zhan K, Zhang A, Liu D (2015) Composition, variation, expression and evolution of low-molecular-weight glutenin subunit genes in *Triticum urartu*. BMC Plant Biol 15:68

Ma ZC, Wei YM, Long H, Yan ZH, Baum B, Zheng YL (2006) Characterization of low-molecular-weight i-type glutenin subunit genes from diploid wheat in relation to the gene family structure. Mol Biol 40:897-906

Martín MA, Martín LM, Alvarez JB (2008) Polymorphisms at the *Gli-*A^u*1* and *Gli-*A^u*2* loci in wild diploid wheat (*Triticum urartu*). Euphytica 163:303-307

Masci S, D'Ovidio R, Lafiandra D, Kasarda DD (1998) Characterization of a low-molecular-weight glutenin subunit gene from bread wheat and the corresponding protein that represents a major subunit of the glutenin polymer. Plant Physiol 118:1147-1158

Masci S, D'Ovidio R, Lafiandra D, Kasarda DD (2000) A 1B-coded low-molecular-weight glutenin subunit associated with quality in durum wheats shows strong similarity to a subunit present in some bread wheat cultivars. Theor Appl Genet 100:396-400

Mason-Gamer RJ, Weil CF, Kellogg EA (1998) Granule-bound starch synthase: structure, function, and phylogenetic utility. Mol Biol Evol 15:1658-1673

Massa AN, Morris CF, Gill BS (2004) Sequence diversity of Puroindoline-a, Puroindoline-b, and the Grain Softness protein genes in *Aegilops tauschii* Coss.

Crop Sci 44:1808-1816

Massa AN, Morris CF (2006) Molecular evolution of the puroindoline-a, puroindoline-b, and grain softness protein-1 genes in the Tribe *Triticeae*. J Mol Evol 63:526-536

Matsuoka Y (2011) Evolution of polyploid *Triticum* wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol 52:750-764

McIntosh RA, Yamazaki Y, Dubcovsky J, Rogers WJ, Morris C, Appels R, Xia XC (2013) Catalogue of gene symbols for wheat. http://www.shigen.nig.ac.jp/wheat/komugi/genes/macgene/2013/ GeneSymbol.pdf. Accessed 20 June 2015

Miller TE (1987) Systematic and evolution. En: Lupton FGH (ed) Wheat breeding: its scientific basis. Chapman & Hall, London. pp 1-30.

Morris C, Rose S (1996) Wheat. En: Henry RJ, Kettlewell PS (eds) Cereal grain quality. Chapman and Hall, New York, pp 3-54

Morris CF (2002) Puroindolines: the molecular genetic basis of wheat grain hardness. Plant Mol Biol 48:633-647

Morris CF, Bhave M (2008) Reconciliation of D-genome puroindoline allele designations with current DNA sequence data. J Cereal Sci 48:277-287

Morris CF, Simeone MC, Gill BS, Mason-Gamer RJ, Lillemo M (2001) Comparison of puroindoline sequences from various diploid members of the *Triticeae* and modern cultivated hexaploid wheats. En: Wootton M, Batey IL, Wrigley CW (eds) Cereals 2000. North Melbourne, Australia: Royal Australian Chemical Institute, pp 678-681.

Morris CF, Geng H, Beecher BS, Ma D (2013) A review of the occurrence of *Grain softness protein-1* genes in wheat (*Triticum aestivum* L.). Plant Mol Biol 83:507-521

Nagy IJ, Takács I, Juhász A, Tamás L, Bedõ Z (2005) Identification of a new class of recombinant prolamin genes in wheat. Genome 48:840-847

Nakamura T, Yamamori M, Hirano H, Hidaka S, Nagamine T (1995) Production of waxy (amylose-free) wheats. Mol Gen Genet 248:253-259

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York.

Ortega R, Alvarez JB, Guzmán C (2014a) Characterization of the *Wx* gene in diploid *Aegilops* species and its potential use in wheat breeding. Genet Resour Crop Evol 61:369-382

Ortega R, Guzmán C, Alvarez JB (2014b) *Wx* gene in diploid wheat: molecular characterization of five novel alleles from einkorn (*Triticum monococcum* L. ssp. *monococcum*) and *T. urartu*. Mol Breed 34:1137-1146

Pauly A, Pareyt B, Fierens E, Delcour JA (2013) Wheat (*Triticum aestivum* L. and *T. turgidum* L. ssp. *durum*) kernel hardness: I. Current view on the role of puroindolines and polar lipids. Compr Rev Food Science Food Saf 12:413-426

Payne P, Law C, Mudd E (1980) Control by homoeologous group 1 chromosomes of the high-molecular-weight subunits of glutenin, a major protein of wheat endosperm. Theor Appl Genet 58:113-120

Payne PI (1987) Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. Ann Rev Plant Physiol 38:141-153

Payne PI, Holt LM, Lawrence GJ, Law CN (1982) The genetics of gliadin and glutenin, the major storage proteins of the wheat endosperm. Plant Foods Hum Nutr 31:229-241

Pei Y-H, Wang A-L, An X-L, Li X-H, Zhang Y-Z, Huang X-Q, Yan Y-M (2007) Characterization and comparative analysis of three low molecular weight glutenin C-subunit genes isolated from *Aegilops tauschii*. Can J Plant Sci 87:273-280

Peña RJ (2002) Wheat for bread and other foods. En: Bread wheat. Improvement and production. Curtis BC, Rajaram S, Gómez Macpherson H (eds.) F.A.O. Rome, pp 483-542

Petersen G, Seberg O, Yde M, Berthelsen K (2006) Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). Mol Phylogenet Evol 39:70-82

Phillips RL, Palombo EA, Panozzo JF, Bhave M (2011) Puroindolines, *Pin* alleles, hordoindolines and grain softness proteins are sources of bactericidal and fungicidal peptides. J Cereal Sci 53:112-117

Pogna NE, Autran JC, Mellini F, Lafiandra D, Feillet P (1990) Chromosome 1B-encoded gliadins and glutenin subunits in durum wheat: Genetics and relationship to gluten strength. J Cereal Sci 11:15-34

Qin LM, Liang Y, Yang DZ, Sun L, Xia GM, Liu SW (2015) Novel LMW glutenin subunit genes from wild emmer wheat (*Triticum turgidum* ssp *dicoccoides*) in relation to *Glu-3* evolution. Dev Genes Evol 225:31-37

Rao VR, Hodgkin T (2002) Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell Tiss Org 68:1-19

Rasheed A, Xia X, Yan Y, Appels R, Mahmood T, He Z (2014) Wheat seed storage proteins: Advances in molecular genetics, diversity and breeding applications. J Cereal Sci 60:11-24

Reynolds NP, Martin JM, Giroux MJ (2010a) Increased wheat grain hardness conferred by novel puroindoline haplotypes from *Aegilops tauschii*. Crop Sci 50:1718-1727

Reynolds NP, Martin JM, Giroux MJ (2010b) Novel *Ha* locus, *Pina-D1c* /*Pinb-D1h*, affects soft wheat milling and baking. Cereal Chem 87:237-242

Riechmann JL, Ito T, Meyerowitz EM (1999) Non-AUG initiation of AGAMOUS mRNA translation in *Arabidopsis thaliana*. Mol Cell Biol 19:8502-8512

Rodríguez-Quijano M, Nieto-Taladriz MT, Carrillo JM (1997) Variation in B-LMW glutenin subunits in Einkorn wheats. Genet Resour Crop Evol 44:539-543

Sarkar P, Stebbins GL (1956) Morphological evidence concerning the origin of the B genome in wheat. Am J Bot 43:297-304

Schneider A, Molnar I, Molnar-Lang M (2008) Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. Euphytica 163:1-19

Sharma HC, Waines JG (1981) The relationships between male and female fertility and among taxa in diploid wheats. Am J Bot 68:449-451

Shewry P (2009) Wheat. J Exp Bot 60:1537-1553

Shewry PR, Halford NG, Belton PS, Tatham AS (2002) The structure and properties of gluten: an elastic protein from wheat grain. Phil Trans R Soc B 357:133-142

Shewry PR, Tatham AS, Barro F, Barceló P, Lazzeri PA (1995) Biotechnology of breadmaking: unraveling and manipulating the multi-protein gluten complex. Bio/Technology 13:1185-1190

Shewry PR, Tatham AS, Forde J, Kreis M, Miflin BJ (1986) The classification and nomenclature of wheat gluten proteins: a reassessment. J Cereal Sci 4:97-106

Simeone MC, Gedye KR, Mason-Gamer R, Gill BS, Morris CF (2006) Conserved regulatory elements identified from a comparative puroindoline gene sequence survey of *Triticum* and *Aegilops* diploid taxa. J Cereal Sci 44:21-33

Singh NK, Shepherd KW (1988) Linkage mapping of genes controlling endosperm proteins in wheat. 1. Genes on the short arms of group-1 chromosomes. Theor Appl Genet 75:628-641

Srivastava JP, Damania AB (1989) Use of collections in cereal improvement in semi-arid areas. En: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 88-104

Stacey J, Isaac P (1994) Isolation of DNA from plants. En: Isaac PG (ed) Methods in molecular biology: protocols for nucleic acid analysis by non-radioactive probes. Humana Press, Totawa, pp 9-15

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 101:11030-11035

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739

Tanaka H, Shimizu R, Tsujimoto H (2005) Genetical analysis of contribution of low-molecular-weight glutenin subunits to dough strength in common wheat (*Triticum aestivum* L.). Euphytica 141:157-162

Terasawa I, Rahman SM, Takata K, Ikeda TM (2012) Distribution of *Hordoindoline* genes in the genus *Hordeum*. Theor Appl Genet 124:143-151

Tranquilli G, Heaton J, Chicaiza O, Dubcovsky J (2002) Substitutions and deletions of genes related to grain hardness in wheat and their effect on grain texture. Crop Sci 42:1812-1817

Turnbull K-M, Rahman S (2002) Endosperm texture in wheat. J Cereal Sci 36:327-337

Vader W, Kooy Y, van Veelen P, de Ru A, Harris D, Benckhuijsen W, Peña S, Mearin L, Drijfhout JW, Koning F (2002) The gluten response in children with celiac disease

is directed toward multiple gliadin and glutenin peptides. Gastroenterology 122:1729-1737

van de Wal Y, Kooy YMC, van Veelen PA, Peña SA, Mearin LM, Molberg Ø, Lundin KEA, Sollid LM, Mutis T, Benckhuijsen WE, Drijfhout JW, Koning F (1998) Small intestinal T cells of celiac disease patients recognize a natural pepsin fragment of gliadin. Proc Natl Acad Sci USA 95:10050-10054

Van den Bulck K, Loosveld A-MA, Courtin CM, Proost P, Van Damme J, Robben J, Mort A, Delcour JA (2002) Amino acid sequence of wheat flour arabinogalactan-peptide, identical to part of grain softness protein GSP-1, leads to improved structural model. Cereal Chem 79:329-331

van Slageren MW (1994) Wild wheats: a monograph of *Aegilops* L. and *Amblyopyrum* (Jaub. & Spach.) Eig. (*Poaceae*). Wageningen Agricultural University Papers, vol. 7

Wang K, Gao L, Wang S, Zhang Y, Li X, Zhang M, Xie Z, Yan Y, Belgard M, Ma W (2011a) Phylogenetic relationship of a new class of LMW-GS genes in the M genome of *Aegilops comosa*. Theor Appl Genet 122:1411-1425

Wang L, Li G, Peña RJ, Xia X, He Z (2010) Development of STS markers and establishment of multiplex PCR for *Glu-A3* alleles in common wheat (*Triticum aestivum* L.). J Cereal Sci 51:305-312

Wang LH, Zhao XL, He ZH, Ma W, Appels R, Peña RJ, Xia XC (2009) Characterization of low-molecular-weight glutenin subunit *Glu-B3* genes and development of STS markers in common wheat (*Triticum aestivum* L.). Theor Appl Genet 118:525-539

Wang S, Li X, Wang K, Wang X, Li S, Zhang Y, Guo G, Zeller FJ, Hsam SLK, Yan Y, Gustafson P (2011b) Phylogenetic analysis of C, M, N, and U genomes and their relationships with *Triticum* and other related genomes as revealed by LMW-GS genes at *Glu-3* loci. Genome 54:273-284

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-76

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu Z-D, Dubcovsky J, Keller B (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell 15:1186-1197

Wilkinson MD, Castells-Brooke N, Shewry PR (2013) Diversity of sequences encoded by the *Gsp-1* genes in wheat and other grass species. J Cereal Sci 57:1-9

Wrigley C, Békés F, Bushuk W (2006) Gliadin and glutenin: the unique balance of wheat quality. AACC International Press, St. Paul

Wu X, Zhao R, Wang D, Bean S, Seib P, Tuinstra M, Campbell M, O'Brien A (2006) Effects of amylose, corn protein, and corn fiber contents on production of ethanol from starch-rich media. Cereal Chem 83:569-575

Xu H, Wang R, Shen X, Zhao Y, Sun G, Zhao H, Guo A (2006) Functional properties of a new low-molecular-weight glutenin subunit gene from a bread wheat cultivar. Theor Appl Genet 113:1295-1303

Yamamori M, Endo T (1996) Variation of starch granule proteins and chromosome mapping of their coding genes in common wheat. Theor Appl Genet 93:275-281

Yanagisawa T, Kiribuchi-Otobe C, Yoshida H (2001) An alanine to threonine change in the Wx-D1 protein reduces GBSS I activity in waxy mutant wheat. Euphytica 21:2009-2014

Yanaka M, Takata K, Terasawa Y, Ikeda TM (2011) Chromosome 5H of *Hordeum* species involved in reduction in grain hardness in wheat genetic background. Theor Appl Genet 123:1013-1018

Yuan ZW, Liu DC, Zhang LQ, Zhang L, Chen WJ, Yan ZH, Zheng YL, Zhang HG, Yen Y (2011) Mitotic illegitimate recombination is a mechanism for novel changes in high-molecular-weight glutenin subunits in wheat-rye hybrids. PloS One 6: e23511

Zaharieva M, Monneveux P (2014) Cultivated einkorn wheat (*Triticum monococcum* L. subsp. *monococcum*): the long life of a founder crop of agriculture. Genet Resour Crop Evol 61:677-706

Zeng M, Morris CF, Batey IL, Wrigley CW (1997) Sources of variation for starch gelatinization, pasting, and gelation properties in wheat. Cereal Chem 74:63-71

Zhang W, Gianibelli MC, Rampling LR, Gale KR (2004) Characterisation and marker development for low molecular weight glutenin genes from *Glu-A3* alleles of bread wheat (*Triticum aestivum* L). Theor Appl Genet 108:1409-1419

Zhang X, Liu D, Jiang W, Guo X, Yang W, Sun J, Ling H, Zhang A (2011) PCR-based isolation and identification of full-length low-molecular-weight glutenin subunit genes in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 123:1293-1305

Zhang X, Liu D, Zhang J, Jiang W, Luo G, Yang W, Sun J, Tong Y, Cui D, Zhang A (2013) Novel insights into the composition, variation, organization, and expression of the low-molecular-weight glutenin subunit gene family in common wheat. J Exp Bot 64:2027-2040

Zhao X, Yang Y, He Z, Lei Z, Ma W, Sun Q, Xia X (2008) Characterization of novel LMW-GS genes at *Glu-D3* locus on chromosome 1D in *Aegilops tauschii*. Hereditas 145:238-250

Zhao XL, Xia XC, He ZH, Gale KR, Lei ZS, Appels R, Ma W (2006) Characterization of three low-molecular-weight *Glu-D3* subunit genes in common wheat. Theor Appl Genet 113:1247-1259

Zhao XL, Xia XC, He ZH, Lei ZS, Appels R, Yang Y, Sun QX, Ma W (2007) Novel DNA variations to characterize low molecular weight glutenin *Glu-D3* genes and develop STS markers in common wheat. Theor Appl Genet 114:451-460

Zohary D, Hopf M (1988) Domestication of plants in the Old World. Oxford Science Publications, Oxford, UK